Tailored IoT & BigData Sandboxes and Testbeds for Smart, Autonomous and Personalized Services in the European Finance and Insurance Services Ecosystem

Infinitech

D3.9 – Automatic Parallelization of Data Streams and Intelligent Pipelining - I

Revision Number	3.0
Task Reference	T3.4
Lead Beneficiary	LXS
Responsible	Ricardo Jiménez-Peris
Partners	LXS, GLA, UNP
Deliverable Type	Report (R)
Dissemination Level	Public (PU)
Due Date	2020-11-30
Delivered Date	2021-05-07
Internal Reviewers	UNIC, ATOS
Quality Assurance	INNOV
Acceptance	WP Leader Accepted and Coordinator Accepted
EC Project Officer	Pierre-Paul Sondag
Programme	HORIZON 2020 - ICT-11-2018
	This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement no 856632

Contributing Partners

Partner Acronym	Role[1]	Author(s)[2]
LXS	Lead Beneficiary	Ricardo Jiménez-Peris
LXS	Contributor	Boyan Kolev, Spencer Pablos, Patricio Martinez, Javier Pereira, Alejandro Ramiro, Jacob Roldan José María Zaragoza, Jesús Manuel Gallego
GLA	Contributor	Richard Mccreadie
UNP	Contributor	Bruno Almeida Tiago Teixeira
UNIC	Internal Reviewer	Marianna Charalampous
ATOS	Internal Reviewer	Jose Gato
INNOV	Quality Assurance	Dimitris Drakoulis

Revision History

Version	Date	Partner(s)	Description
0.1	2021-04-01	LXS	ToC Version and updated initial input of D3.9
0.2	2021-04-29	LXS, UNP	Input on sections 2, 3, 4
0.3	2021-05-03	GLA	Input on section 5
0.4	2021-05-04	LXS, GLA	Finalizing the document
1.0	2021-04-05	LXS	Submitted for internal review
1.1	2021-05-05	UNIC	Internal review
1.2	2021-05-05	ATOS	Internal review
2.0	2021-05-05	LXS	Submitted for internal QA
3.0	2021-05-07	INNOV, LXS	Document finalized

[1] Lead Beneficiary, Contributor, Internal Reviewer, Quality Assurance

[2] Can be left void

Executive Summary

The goal of task T3.4 "Automated Parallelization of Data Streams and Intelligent Data Pipelining" is to firstly provide the INFINITECH approach for intelligent data pipelines, and secondly to allow for the automation of the deployment of parallelized data streams. Modern applications being currently used by data-driven organizations, such as those belonging to the finance and insurance sector, requires to process data streams along with data persistently stored in a data base. A key requirement for such organizations is that the processing must take place in real-time providing real-time results, alerts or notifications in order for instance to detect fraud finance transactions the moment they are being occurred, detect possible indications for money laundering or provide real-time risk assessment among other needs. Towards this direction, streaming processing frameworks have been used during the last decade in order to process streaming data coming from various sources, in combination with data persistently stored in a database that can be considered as data *at-rest.* However, processing data *at-rest* introduces an inherit significant latency, as data access involves expensive I/O operations, which are not suitable for streaming processing. Due to this, various architectural designs have been proposed and are used in the modern landscape that deals with such problems. They tend to formulate data pipelines, moving data from different sources to other data management systems, in order to allow for an efficient processing in real-time. However, they are far from being considered as *intelligent*, and each of the proposed approaches comes with their own barriers and drawbacks.

A second key requirement for data-driven organizations in finance and insurance sector is to be able to cope with diverse workloads and continue provide results in real-time even when there is a burst of incoming data load from a stream. This might happen in case of having a stream consuming data feeds from social media in order to perform a sentiment analysis and an important event or incident takes place which will make the social community to response by posting an increased number of tweets or articles. Another example is the unexpected currency devaluation that will most likely trigger numerous of finance transactions with people and organizations change their currencies. The problem with the current landscape is that modern streaming processing frameworks allow for static deployments of data streams that consist of several operators, in order to serve an expected input workload. In case of such scenarios, an unexpected burst of the incoming workload might saturate the resources devoted for the initial deployment which cannot provide the results in real-time, or even worse, might be crashed.

In order to cope with those two requirements and overcome the current barriers of the modern landscape, we envision the INFINITECH approach for Intelligent Data Pipelines and the parallelized data stream processing using Apache Flink as the baseline technology for the INFINITECH streaming processing framework along with the Kubernetes. In our solution we provide a holistic approach for data pipelines that makes use of the key innovations and technologies provided by INFINITECH and implemented in the rest of the tasks of the project related with the data management activities. Our solution solves all problems with the different types of storage, use of different types of databases for persistently store data and allowing for efficient query processing, handling aggregates and dealing with snapshots of data. Moreover, we have designed our solution for parallelized data stream processing, allowing for the deployed operators to save and restore their state and allowing for online reconfiguration of the Flink clusters, which enables elastic scalability by programmatically scaling the clusters.

This document reports on the work that has been done towards those two main objectives, at this phase of the project. An initial state-of-the-art analysis of the current status of the data pipelines and data stream parallelization has been provided along with their current barriers problems and solutions used so far. Then, we provide the current landscape of data pipelining in modern enterprises today, we analyzed the different architectures used currently to cope with the inherit problems of combining streaming data with data *at-rest*, and what are their benefits and drawbacks of each solution, and then we defend on how the INFINITECH approach can be used to solve those issues in an holistic manner, taking advantage and using the developments of the other tasks and the INFINISTORE as its basic pillars. Then, we describe the initial

design of our solution to allow the dynamic redeployment of the parallelized data streams to enable the elastic scalability of the deployed operators.

In this first version of the deliverable of task T3.4 "Automated Parallelization of Data Streams and Intelligent Data Pipelining" we introduced our vision and design of the overall solution, as the main focus at this phase of the project was to develop the baseline technologies that are needed and will be used as the basic pillars of our overall implementation. In the second version of this report, a more analytical description will provide the implementation details and experimentation results.

Table of Contents

List of Figures

Abbreviations/Acronyms

ACID	Atomicity, Consistency, Isolation, Durability
API	Application Programming Interface
CDC	Change Data Capture
CEP	Complex Event Processing
CPU	Central Processing Unit
DoA	Description of Action
DNS	Domain Name System
ELT	Extract, Load, Transform
ETL	Extract, Transform, Load
FP7	7th Framework Program
HDFS	Hadoop Distributed FileSystem
HTAP	Hybrid Transactional and Analytical Processing
I/O	Input / Output
INFINISTORE	The INFINITECH data management layer based on the LeanXcale database
IoT	Internet of Things
JSON	Javascript Object Notation
MoM	Monitor of Monitors
OLAP	Online Analytical Processing
SQL	Structured Query Language
WP	Work Package
XML	Extensible Markup Language

1 Introduction

Data-driven organizations nowadays are increasingly need the combination of streaming data with persistently stored data, usually called data *in-flight* and data *at-rest*. Streaming processing frameworks that have been adopted and widely used during the last decade are being enhanced with the additional functionalities for solving the inherit barriers that database management systems introduce to an integrated solution. The most important obstacles are the latency of an analytical query processing operation over a persistent data store, and that datastores are not designed to support data ingestion in very high rates in order to serve in increased data load coming from a stream. Due to this, they tend to use a variety of different data management systems, from operational datastores, to data warehouses and data lakes. Operational datastores allows to persistently write data, but they are not appropriate to perform analytical query processing over bigdata. For this, data is being periodically moved to data warehouses, that are designed to only allow for read operations, using sophisticated indices and data structures that can boost the performance of such operations. Data lakes can be also used as a cheap solution to store historical data that does not require frequent processing. Operational datastore on the other hand can be further divided to different categories: traditional SQL datastores that ensures consistency in terms of database transactions, which are critical for applications in the insurance and finance sector. However, they are not designed to scale out, and they are inefficient in performing analytics on the same time. Due to this, other types of operational datastores have been adapted during the last decade, commonly refered as NoSQL datastores (or their evolutions which are widely known as NewSQL), which sacrifices transactional semantics for the sake of scalability. Usually they can scale out more efficiently and can serve ingestions in very high rates. However, they lack of rich query processing capabilities. As a result, modern enterprises use a variety of different datastores, stream data managers and tools such as Kafka and machine learning infrastructure. This makes the data pipelines more complex and problematic, as they need to move data across the different databases, in order take advantage of the benefits of each one. For that, they rely on expensive (ETLs (Extract, Transform, Load) and they perform periodic batch processing, which is not suitable when there is the need for real-time data analytics. Periodic batch processing makes the data used in an processing framework obsolete, as ETLs take place periodically, once every day or during weekends.

In order to overcome these problems, different architecture designs have been proposed and adapted in modern enterprises. For instance, lambda architectures have been widely adopted to solve the problems of complexity mixing different databases and the need for real-time processing. But they are very complex, they consist of different layers, with different codebase, while their maintenance is hard to keep. Other architectures rely on moving data from operational to analytical datastores and vice-versa, using architectural designs such as *current-historical data splitting*, *data warehouse or operational data offloading* and *database sharding*. All of these come with the drawback that query processing takes place over a snapshot of the dataset, and the results are obsolete. Other approaches aim at improving the latency of the execution of an analytical query, which involves aggregations. This is crucial as the response time must be very low in order to be used with combination with streaming operators. *Detail-aggregate view splitting, in-memory application aggregations* and *federated aggregations* are techniques widely used to solve these issues, with the drawback of sacrificing the consistency and accuracy of the results. We provide details on these designs in section 3 of this report.

To solve these issues, we propose the INFINITECH approach for Intelligent Data pipelines, that provides an holistic solution for data pipelines, solving major problems with different types of storage, handling of aggregates and deal with snapshot databases. Our proposed analytical pipeline will address all of the above identified architectural patterns for data pipelining, combining data streaming and data at rest, taking advantage of the technologies and innovations developed in INFINITECH that break through the current barriers of modern applications in finance and insurance sectors, and have been reported so far in the corresponding deliverables such as D3.1, D3.4, D3.6, D5.1 and D5.4 that report the work that has been carried out so far in the rest of the tasks related with the data management layer of IFINITECH. We will build our solution taking these prototypes as the basic pillars.

Another key requirement for modern data-driven organizations using streaming processing frameworks is the ability to cope with diverse incoming workloads. So far, current solutions allow only for static deployments of data stream operators. This is problematic in the sense that in cases of unexpected peaks of incoming data coming from a stream, the static deployment cannot scale out to cope with that need.

In order to help solving the previous explained issues and according to the different INFINITECH pilots, we propose a solution based on Apache Flink as the streaming processing framework of INFINITECH, integrated with the INFINISTORE as the data management layer and in combination with the Kubernetes as the container-orchestration framework. By using INFINISTORE, we can cope with the majority of technology challenges for data pipelines that require complex architectures introducing additional barriers, as it will be explained in section 4. The use of Flink allows us to also extend their operators to store and restore their state using checkpoints. By doing so, we are now capable of redeploying a Flink cluster, increasing the number of instances of the operators, and restore their relevant state. This allows programmatically scaling the Flink clusters and providing dynamic redeployments in order to cope with these diverse workloads.

1.1. Objective of the Deliverable

The objective of this deliverable is to report the work that has been done in the context of the task T3.4 “Automated Parallelization of Data Streams and Intelligent Data Pipelining” at this phase of the project. This task lasts until M30, and therefore, two additional versions will be released, extending and modifying the content of this document. During this phase, the identification of the problem that needs to be solved took place, by performing a throughtfull state-of-the-art analysis on problems and solutions of data pipelines and data stream parallelization, along with a detailed analysis of their current landscape in modern enterprises of today. We identify which different technologies are mostly used in order to combine processing of streaming data with data at-rest in a data pipeline, and the dominant architectural designs used so far, in order to identify the current drawbacks. We then propose the INFINITECH approach for data pipelines, that will solve those problems in an holistic manner. Moreover, we designed our solution that allows for dynamic redeployments of streaming operators. This will allow for a dynamic scaling of those operators, which is a current challenge.

In the next iterations of this deliverable, a more detailed description of our implementation will take place, along with the experimentation with use case scenarios coming from the pilots of INFINITECH. However, our solution is case agnostic, and has been designed to fit any other user story.

1.2. Insights from other Tasks and Deliverables

The work that is reported in this deliverable is based on the overview of baseline technologies defined in WP2. This task is based on the technologies and innovations implemented under the scope of the tasks of the project that are related with the data management layer and make use of those at the basic pillars. As a result, is very closely related with T3.1 “Framework for Seamless Data Management and HTAP”, which provides the fundamentals that allows the hybrid transactional and analytical processing, allowing the for query processing over live data added in the operational datastore, removing the need to migrate data to a data warehouse. Moreover, T3.1 allows for the direct data ingestion in very high rates, which removes this barrier from our solution. T3.2 “Polyglot Persistance over BigData, IoT and Open Data Sources” provides the polyglot extensions in the level of INFINISTORE query engine that allows for a unified manner to query data that are stored in a different data sources. T5.3 “Declarative Real-Time Data Analytics” implements the *online aggregates* that will be massively used in our solution, removing the inherit barriers of data consistency that various architectural designs suffers when trying to pre-calculate the results of the analytical operations in order to boost the performance of such operations. Finally, T3.3 “Integrated Querying of Streaming Data and Data at Rest” provides the Apache Flink as the streaming query processing framework of INFINITECH, integrated with the data management layer (the INFINISTORE) to allow the combination of streaming processing with data *at-rest* in INFINISTORE. With this integration, we are now capable of taking advantage of all these aforementioned technologies into our INFINITECH approach for Intelligent Data Pipelines. Last but not least, the INFINITECH way for deployment, using tailored sandboxes that rely on the Kubernetes, as implemented under the scope of WP6 "Tailored Sandboxes and Testbeds for Experimentation and Validation" and the provision of the reference testbed, in combination with the use of Apache Flink as the baseline technology for streaming processing, allows for the parallelization of the data streams which is the second important objective of the task.

1.3. Structure

This document is structured as follows: Section 1 introduces the document, putting the work reported in this deliverable under the context of the project, highlighting its relation with the tasks related with the data management activities of the project. Section 2 provides an extensive state-of-the-art analysis on data stream parallelization, describing the problems and the most adapted solutions. Section 3 describes the current landscape of modern applications that make use of the data pipelines while section 4 introduces our vision for the INFINITECH approach for Intelligent Data Pipelines, after providing an analytical survey of modern architectural designs along with their drawbacks and inherit issues, describing how our approach can provide a holistic way to solve all those issues. Section 5 deals with the problem of dynamic reconfiguration of streaming operators that will allow for dynamic scalability and describes the design of our solution. Finally, section 6 concludes the document and describes the next steps towards the delivery of our prototype.

2 State-of-the-Art Analysis on Data Stream Parallelization

2.1 Introduction

There is an increasing demand in data-driven organizations to process data streams as opposed to only stored data. A data stream is a sequence of tuples with some pre-defined schema. The main difference with a database is that a database query processes a snapshot of data at a particular point in time and produces the answer, while a streaming query is continuous and produces results continuously. Basically, a stream comes from a data source and contains tuples. A data streaming query processes this continuous sequence of tuples producing a continuous stream of results. Data streaming has many applications in the financial and insurance world such as fraud detection, IoT, stock trading, etc. For this reason, they are becoming more and more important in data-driven organizations.

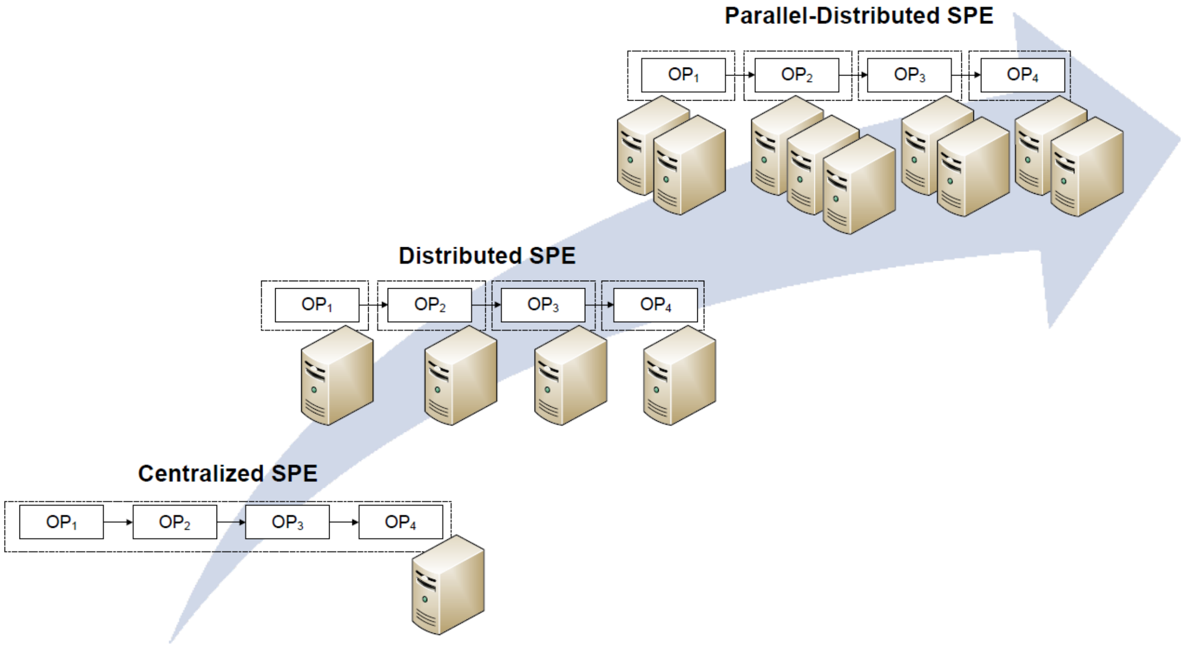

Figure 1: Stream Processing Engines Evolution

Data streaming engines started as centralized systems such as NiagaraCQ (J. Chen, 2000) and TelegraphCQ (S. Chandrasekaran, 2003), or Aurora (D. Carney, 2002). The main issue with centralized engines is that they are limited to the capacity of a single node and they cannot process large stream volumes. For this reason, distributed stream processing systems started to be built introducing different kinds of parallelism in the query processing. In the next section we elaborate on the different ways to introduce parallelism in query processing to enable distributed stream processing.

2.2 Query Parallelism & Data Partitioning

There are two broad approaches of introducing parallelism in query processing (Valduriez, 2020):

1. **Inter-query parallelism.** This parallelism basically enables to run different queries in parallel. Although useful if there are many queries, a not very common case in data streaming, it doesn't help to process large volume streams, that is, streams with a high rate of tuples.

2. **Intra-query parallelism.** This parallelism actually enables to accelerate even a single query. It basically consists in parallelism the processing within the query. There are two mechanisms that can be used:

    a. **Inter-operator parallelism.** This parallelism lies in having multiple operators within the query plan to run in parallel and process tuples at the same time. It can be helpful if there are many operators, that again, it is not the common case in data streaming systems.

    b. **Intra-operator parallelism.** This type of parallelism parallelizes the processing of a single operator within the query plan, enabling multiple threads/processes to process different subsets of the incoming data streaming to the operator.

Centralized approaches already provide inter-query parallelism. One can run multiple queries on the same stream engine, and one could scale a little bit by using multiple centralized engines, one for each different query. Intra-query parallelism came with distributed data streaming frameworks. Some of the pioneering ones are Borealis (Daniel J. Abadi, 2005) and StreamCloud (V. Gulisano, 2010). Distributed data streaming systems introduced intra-query parallelism. Borealis introduced inter-query parallelism via inter-operator parallelism. While StreamCloud introduced intra-operator parallelism, thus enabling scale out to large volume data streams. This intra-operator parallelism pioneered by StreamCloud has become the standard in modern data streaming systems such as Flink (P. Carbone, 2015) (originally called stratosphere (A. Alexandrov, 2014)) and Spark streaming (Matei Zaharia, 2013). StreamCloud was the main outcome of the Stream FP7 project and has become one of the main references in distributed data streaming. It was also the first data streaming system to implement elasticity (V. Gulisano R. J.-M., 2012). This evolution of data streaming engines is depicted in **Error! Reference source not found.**.

2.3 Data Partitioning

Another aspect related to the query parallelism and needed to make such systems work is data partitioning. Once there is intra-operator parallelism, the question is how to partition the data across the different instances of a given operator. There are two big approaches for partitioning the data, vertical and horizontal. Vertical partitioning lies in splitting the data as different subsets of columns. However, data streaming queries are not very amenable for vertical partitioning in general. Horizontal partitioning splits data as disjoint sets of full rows/events. Horizontal partitioning can be further classified into range, hash partitioning, and round-robin. Range partitioning relies on each partition being a range of the distribution key. With hash partitioning the distribution key is hashed and the modulus is obtained dividing by a number of buckets. The result is a bucket number that is used as the distribution unit. In round-robin events are sent in a round-robin fashion to the parallel operators.

However, it should be noted that data partitioning depends on which category of operator is being applied. There are stateless and stateful operators. Stateless operators perform their function independently of any previous input. An example of stateless operators is a filter. Stateful operators on the other hand perform a computation over several rows/events. An example would be an aggregation.

Range partitioning has the disadvantage that it requires tuning to guarantee a good load balancing across parallel instances of the operator. Hashing on the other hand provides good load balancing by default, although it has an extra processing cost to compute the hash over the distribution key. Round-robin partitioning only works for stateless operators. If the partitioning is not done right, the parallelism does not improve the performance, simply wasting hardware.

2.4 Genuine Stream Processing vs. Batch Processing

Another dimension in the comparison of streaming engines is whether they are genuine stream processing engines or they process batches. Genuine stream processing engines, such as StreamCloud or Flink, process events as they are produced yielding real-time processing. However, batch-based engines such as Spark streaming have some delay in the processing of events (e.g a few seconds), since they aim to buffer a set of events before processing begins. Processing in batches can be advantageous for more efficient processing. However, it introduces a delay of seconds that can be detrimental for event-oriented applications.

2.5 Fault Tolerance and Message Processing Coherence Guarantees

Another aspect of distributed stream engines is whether they provide a mechanism to attain high availability in the advent of failures or not, and in affirmative case what are the message processing coherence guarantees.

There are different fault tolerance techniques that can be used in data streaming to provide high availability of different degrees:

Active Replication:

> It lies on having each operator instance running on two or more nodes and then all instances will receive the same data and produce the same output. This requires sending data from a replicated operator to another replicated operator and guarantee 1-copy semantics, that is, guaranteeing that the replicated data streaming engine has the same functional behaviour as the non-replicated one.

Checkpointing:

> Under a checkpointing approach, the state of stream processing pipeline components are periodically saved to a persistant storage medium. During operation, the stream processing platform tracks what parts of the stream have finished processing. If a failure occurs, then the pipeline (either as a whole or as only a subset of failed components) are rolled-back to the last checkpoint and processing continues from the start of that checkpoint.

2.6 Distributed Data Streaming Engine Components

The architecture of distributed data streaming engines shares some similarities from a functional point of view. They have, in general, the following layers:

1. Data ingestion layer.

   This layer is specialized to capture data from different data sources. These data sources can be monitoring probes or logs, information systems, IoT devices, etc. Basically, the different system specializes in being able to capture this data with minimal effort providing ways to automatically extract the data from the sources or supporting a data format that the data sources use such as text, JSON, binary, etc. They can also use a generic mechanism to connect to the data sources such as sockets or REST interfaces.

2. Data processing layer.

   This layer is in charge of actually processing the continuous queries. One of the most common ways is that the data processing layer is represented as a set of containers where one can deploy one or more data streaming operators. Data streaming operators are typically algebraic operators able to do basic functions such as filtering with a predicate, doing a vertical projection (i.e., selecting a subset of the columns), doing an aggregation (e.g., the sum of some column), or even joining two data streams (e.g., join together tuples with the same key).

3. Storage layer.

   Although data streaming engines are typically managing in memory data, many of them provide interfaces to storage of different kinds. It can be from file systems such as HDFS, to key-value data stores such as HBase or Cassandra or even full-fledged relational SQL databases such as PosgreSQL.

4. Output layer

   The output layer provides the interface to send the output of the continuous streaming queries running on the streaming engine. The output can be dashboards, visualization tools, a file, or even a socket to connect to an arbitrary application.

5. Management layer

   This is the layer handling the deployment and decommissioning of queries, fault-tolerance, etc. It orchestrates the different nodes used for the distributed data streaming engine to process a set of continuous data streaming queries and connect them to the data sources and produce the output data to the data sinks.

2.7 Distributed Data Streaming Engine Categories

Distributed data streaming engines have undergone specialization and the following different categories can be identified:

1. General purpose data streaming frameworks.

   They aim at providing a framework where continuous data streaming queries can be deployed and processed at different scales. They allow to express queries either as an acyclic directed graph of algebraic operators or in a query language. Early data streaming engines, such as Borealis and StreamCloud, basically they allow to process queries, while modern frameworks such as Flink and Spark Streaming they provide other functionalities needed by enterprises. For instance, Spark Streaming is integrated with Spark for doing machine learning.

2. Complex Event Processing (CEP).

   They provide support to write business rules to deal with the identification of particular events from continuous streams of information, such as a threat situation in security, when to buy or sell stocks in stock trading, etc. These rules enable to detect event patterns, abstract events or event-driven processes, model event hierarchies, detecting event relationships (causality, membership, timing), and do similar processing as with data streaming systems such as filtering, aggregation, and transformation. Examples of CEP systems are Esper (Espertech, s.f.) and StreamBase (Tibco, s.f.).

3. Online machine learning systems.

   These systems apply machine learning over data streams to provide a continuous learning over the data stream. This is interesting when one cannot train over a full dataset or new patterns can appear over time. One sample system in this category is SAMOA (Scalable Advanced Massive Online Analysis) (Gianmarco De Francisci Morales, 2015). SAMOA actually can work over different data streaming engines such as Storm (http://storm.apache.org/, s.f.), S4 (http://incubator.apache.org/s4, s.f.), and Samza (http://samza.incubator.apache.org, s.f.).

4. Streaming graph analytics.

   They basically keep a graph in memory updated by means of streaming actions and provide real-time processing over the graph such as recommendations, etc. The examples on this area mainly come from Twitter such as GraphJet (A. Sharma, 2016).

2.8 Window Programming Models

Data streaming engines work on the idea of an infinite stream of events. It is equivalent to a regular database table with infinite rows. The processing is made in chunks. Actually, a sliding window over this stream of events. Windows can be defined based on time or number of events. The nature of the window sliding can be different and basically, there are three main window programming models:

1. Fixed or tumbling windows.

These windows split time in consecutive intervals. Events are considered in the interval their timestamp belongs.

2. Sliding windows.

   Sliding windows are more generic. They support overlapping windows. They are defined by two parameters: length of the window and slide. The length indicates how long is the interval. The slide how much is shifted the window during each step. If the length is 10 and the shift 5, there will be windows from 1 to 10, 5 to 15, 11 to 20, etc. If the shift is equal to the length, then they behave like fixed windows.

3. Session windows.

   Sessions are defined by certain thresholds, typically certain time of inactivity. They are used for user-oriented input in which users are active and then after they are inactive for some time the session is considered to be finished. When activity restarts it starts in a new session.

2.9 Data Source Interaction Models

There are basically two modes of interaction with data sources: push and pull. In the push mode, the data source sends data through an API as soon as it has new data. In the pull mode, there is an agent at the data source side that periodically checks whether there is data available and sends it when new data is found. The push mode gives the best response times because data is sent as soon as it is available. Also the pull mode has the shortcoming that the frequency of pulling has to be higher than the frequency of data generation, since otherwise it leads to data loss.

3 The Landscape of Data Pipelining at Enterprises Today

In the current data management landscape, there are clearly two big families of data management, streaming data and data at rest. The latter is the most extended, however, the former is gaining traction to solve problems not amenable for the latter. There are some key differences between data streaming and persistent data stores. The first difference is the fact that data streaming queries are continuous and work over sliding windows, while persistent databases perform point-in-time queries executed just once over stored data. The second difference is that data streaming engines rely on in memory state that allow them to process efficiently large volumes of streaming data while persistent databases have to access persisted data that is far more costly and is what makes them slower and not being able to process the same volume of streaming data per node.

However, each family has a number of possibilities, especially, the one related to persistent databases. Let us first have a look at this landscape to better understand how INFINITECH can help in the problem of data pipelining.

Persistent databases can be classified first into:

1. Operational databases.

   These databases store data in persistent media. They allow to update the data while the data is being read. The consistency guarantees that are given with concurrent reads and writes vary. Operational databases, because they can be used for mission critical applications, might provide capabilities for attaining high availability that tolerates node failures and in some cases they can even tolerate data centre disasters leading to the whole loss or lack of availability of a whole data centre. The source of these disasters can be from a natural disaster like a flood, a fire, the loss of electric power, the loss of Internet connectivity, a Distributed Denial of Service attack resulting in the loss of CPU power and/or network bandwidth, or the saturation of some critical resource like DNS, etc.

2. Data warehouses.

   Data warehouses are informational databases. They are designed only to query data after ingesting it. They do not allow modifications, simply loading the data, and after the loading is complete, querying the data. They specialize on speeding up the queries by means of OLAP (On Line Analytical Processing) capabilities. OLAP capabilities are attained by introducing intra-query parallelism typically in the form of intra-operator parallelism. They typically use a customised storage model to accelerate the analytical queries by using a columnar model or they use an in-memory architecture.

3. Data lakes.

   They are used as scalable cheap storage where to keep historical data at affordable prices. The motivation of keeping this historical data might be legal requirements of data retention but more recently the motivation is from the business side to have enough data to be able to train machine learning models in a more effective way by reaching a critical mass of data in terms of time, but also in terms of detail. Some organizations use data lakes as cheap data warehouses when the queries are not especially demanding in terms of efficiency. A data lake might require more than an order of magnitude higher resources for an analytical query with a target response time than a data warehouse, while the price follows an inverse relationship.

Operational databases can themselves be classified in three broad categories:

1. Traditional SQL databases.

   Traditional SQL operational databases are characterized by two facts. The first one is that they provide SQL as query language. The second one is that they provide the so-called ACID guarantees over the data. We discuss later these ACID properties in detail. The main limitations of traditional SQL databases is their scalability, typically they either don't scale out or scale out logarithmically that means that their cost grows exponentially with the scale of the workload to be processed. They typically provide mechanisms for high availability that guarantee the ACID properties what is technically known as 1-copy consistency guarantees. The second limitation that they have is that they ingest data very inefficiently so they are not able to insert or update data at high rates. Their lack of linear scalability also results in exponentially growth of cost.

2. No-SQL databases

   No-SQL databases is a category with a number of different kinds of databases that are characterized by addressing requirements not well-addressed by traditional SQL databases. There are four main kinds of No-SQL databases as we will see later. Basically, they address the lack of flexibility of the relational schema that is very rigid and forces to know in advance all the fields of each row in the database and they are very disruptive when this schema has to be changed, typically resulting in having the database or at least the involved tables not available during the schema change. No-SQL databases fail to provide ACID consistency guarantees. On the other hand, most of them they are able to scale out, although not all kinds have this ability. Some of them are able to scale out but not linearly or not to large numbers of nodes.

3. NewSQL databases

   NewSQL databases appear as a new approach to address the requirements of SQL databases but trying to remove part or all of their limitations. The direction of NewSQL databases lie in bringing new capabilities to old traditional SQL databases by leveraging approaches from NoSQL and/or new data warehouse technologies. Some try to improve the scalability of storage. That is normally achieved by relying on some NoSQL technology or adopting an approach similar to some NoSQL technology. Scaling queries was an already solved problem. However, scaling inserts and updates had two problems. The first one is the inefficiency of ingesting data. The second one is that inability to scale out to large scale the ACID properties, that is, transactional management. Others have tried to overcome the lack of scalability of the ingestion while others the lack of scalability of transactional management.

NoSQL databases have different flavours and typically are divided into four categories:

1. Key-value data stores.

   They are schema-less and allow any value associated to a key. In most cases they attain linear scalability. Basically, each instance processes a fraction of the load. Since operations are based on an individual key-value pair, the scalability does not pose any challenge and most of the times is achieved. The schema-less approach provides a lot of flexibility. Basically, each row can have a totally different schema. Obviously, that is not how they are used. But they allow to evolve the schema without any major disruption. Of course, the queries have to do the extra work of being to understand rows with different schema versions, but since normally, the schema are additive, they add new columns or new variants, it is easy to handle. Key-value data stores excel at ingesting data very efficiently. Due to the fact that they are schema-less they can just store the data. This is very inefficient for querying, and normally they provide very little capabilities for querying such as getting the value associated to a key. In most cases they are based on hashing so they are unable to perform basic range scans. Example of key-value data stores are Cassandra and DynamoDB.

2. Document oriented databases

   They support semi structured data normally written in a language such as JSON or XML. Their main capability is that being able to store data in one of these languages efficiently and being able to perform queries for these data in an effective way. Representing these data in SQL is just a nightmare and doing queries of this relational schema even a worse nightmare. That is why they have succeeded. Some of them scale out in a limited way and not linearly, whilst some others do better and scale out linearly. The main shortcoming is that they do not support the ACID properties and that they are inefficient querying data that is structured in nature. Structured data can be queried one to two orders of magnitude more efficiently with SQL databases. Examples in this category are MongoDB and Couchbase.

3. Graph databases

   They specialize on storing and querying graph data. Graph data represented in a relational format becomes very expensive to query. The reason is that to traverse a path from a given vertex in the graph, one has to perform many queries, one per edge stemming from the vertex and as many times as the longest path sought in the graph. These results into too many client-server invocations. If the graph does not fit into memory, then it is even a bigger disaster since disk accesses will be involved for most of the queries. Also, the queries cannot be programmed in SQL and has to be performed programmatically. Graph databases on the other hand, they have a query language in which with a single invocation solve the problem. Data is stored to maximize locality of a vertex with contiguous vertexes. However, graph databases when they don’t fit in a single node, then they start suffering from the same problem when they become distributed losing their efficiency and having a performance gain that is lost very soon with the system size in number of nodes. At some point a relational schema solution becomes more efficient than the graph solution for a large number of nodes. A widely used graph database is Neo4J.

4. Wide column data stores

   These data stores have more capabilities than key-value data stores. They typically perform range partitioning thus, supporting range queries. In fact, they might support some limited basic filtering. They are still schemaless. They also support vertical partitioning that can be convenient when the number of columns is very high. They have some notion of schema but still they are quite flexible in it. Example of this kind of data stores are BigTable and HBase.

In addition to the above categories, we have stream data managers already explored in the previous section and systems like Kafka that are streaming data managers but are persistent, and machine learning infrastructure such as Map Reduce, Spark and Pandas. Large organizations such as the ones in the Finance and Insurance verticals, typically have databases of many of the above kinds, with many instances of each category using the same brand or even different brands.

The main issue is that for analytical pipelines they have to move data across databases, many times having to adapt, modify or enrich the data. For this, ETL (Extract Transform Load) tools have been being used to perform batch processing and moving data from one database into another transforming the data as necessary. More recently a different approach such as ELT (Extract Load Transform) has been used. However, batch processing is not always feasible and it is required to acquire the data in real-time or near real-time when it is updated. For this purpose, CDC (Change Data Capture) infrastructure has been created that enables to get the changes from an operational database and then do something with these changes like storing it in some other database or do some processing like triggering events. In this task, we envision to take benefit from the CDC infrastructure in order to create the Intelligent Data Pipeline that INFINITECH needs to provide.

4 Intelligent Data Pipeline: The INFINITECH Approach

In INFINITECH, we are looking at how to simplify these data pipelines and adopt a uniform simple approach for them. Data pipelines get complicated mainly due to the mismatch of capabilities across the different kinds of systems. Many times data pipelines get very complex because of real-time requirements. One solultion is the adaptaion of an architecture, which is well known as **lambda architecture**.

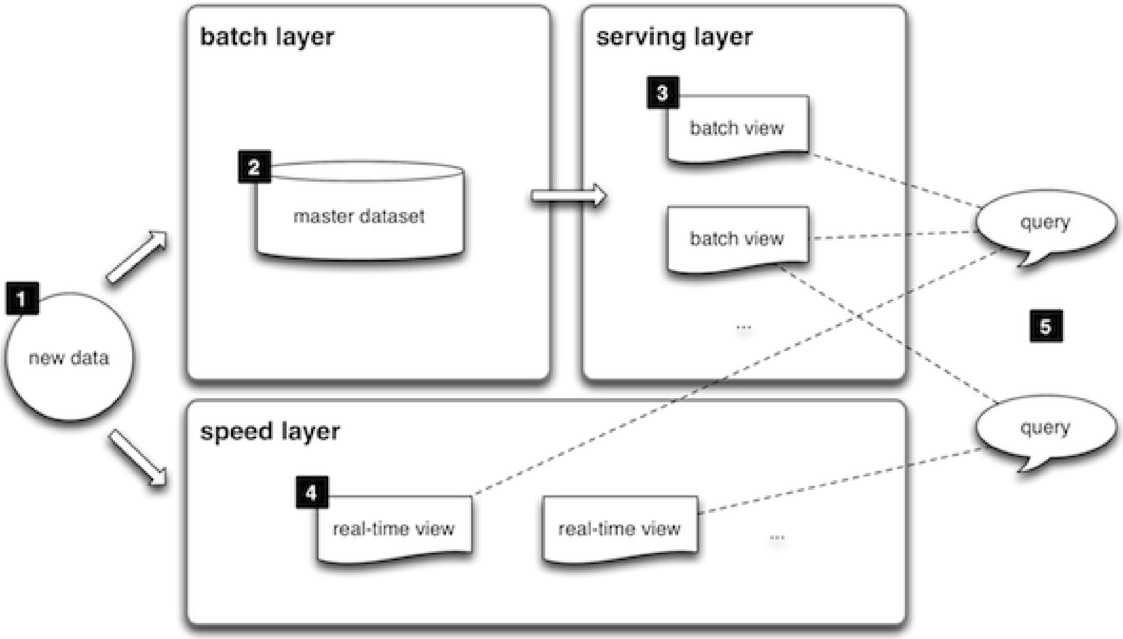

Figure 2: A typical lambda architecture

The lambda architecture combines techniques from batch processing with data streaming to be able to process data in a real-time manner. The lambda architecture is motivated by the lack of scalability of operational SQL databases. The architecture consists of three layers:

1. Batch layer.

   It is based on append only storage, typically a data lake, such as HDFS. Then, it relies on map-reduce for processing new batches of data in the forms of files. This batch layer provides a view in a read-only database. Depending on the problem being solved, the output might need to fully re-compute all the data to be accurate. After each iteration, a new view of the current data is provided. This approach is quite inefficient but it is solving a scalability problem that when it was invented did not have a good solution, the processing of tweets in Twitter.

2. Speed layer.

   This layer is based on data streaming. In the original system at twitter it was accomplished by the Storm data streaming engine. It basically processes new data to complement the batch view with the most recent data. This layer does not aim for accuracy, which is usually very crucial for applications in the insurance and finance sector, but it provides more recent data to the global view achieved with the architecture.

3. Serving layer.

> The serving layer processes the queries over the views provided by both the batch and speed layer. Batch views are indexed to be able to answer queries with low response times and combines them with the real-time view to provide the answer to the query combining both real-time data and historical data. This layer typically uses some key-value data store to implement the indexes over the batch views.

The main shortcoming of the lambda architectures is its complexity and the need to have totally different code bases for each layer that have to be coordinated to be fully in sync. Maintenance of the platform is very hard since debugging implies understanding the different layers with totally different natures, involved technologies and approaches.

Other more traditional architectures are based on combining an operational database with a data warehouse. The operational database deals with more recent data while the data warehouse deals with historical data. In this architecture, queries can only see either the recent data or historical data, but not a combination of both as it was done in the lambda architecture. In this architecture there is a periodic process that copies data from the operational database into the data warehouse. This periodic process has to be performed very carefully since it can hamper the quality of service of the operational database. This periodic process is most of the time achieved by ETL tools. Many times this process is performed over the weekends in businesses where their main workload comes during weekdays. Another problem that this architecture exhibits is the fact that the data warehouse typically cannot be queried while it is being loaded, at least the tables that are being loaded. This forces to split the time of the data warehouse into loading and processing. When the loading process is daily, finally the day is split into loading and processing. The processing time consumes a fraction of hours of the day that depends on the analytical queries that have to be answered daily. It leaves a window of time for loading data that is the remaining hours of the day. At some point data warehouses cannot ingest more data because the loading window is exhausted. We name this architecture **current-historical data splitting**.

Due to the saturation of the data warehouse is a common problem, another architectural pattern has been devised to deal with this issue in an architectural pattern that we name **data warehouse offloading**. This pattern relies in creating small views of the data contained by the data warehouse and store them on independent databases, typically called **data marts**. Depending on the size of the data and the complexity of the queries data marts can be handled by operational SQL databases or they might need a data manager with OLAP capabilities that might be another data warehouse or a data lake plus an OLAP engine that works over data lakes.

In some other cases, the problem lies in the fact that the operational database cannot handle the whole workload due to its lack of scalability and part of this workload can be performed without being real-time. In these cases, a copy of the database or the relevant part of the data of the database is copied into another operational database during the time that the operator is not being used, normally, weekends or nights, depending on how long the copy of the database takes. If it takes less than the night time is performed daily. If it takes more than that time is performed over the weekend. If it takes more than then weekend, it cannot be done with this architectural pattern. We call this architectural pattern **database snapshotting.**

In other cases, there are real-time or quasi real-time requirements and the database snapshotting does not solve the problem. In this case, a CDC (Change Data Capture) system is used that captures changes in the operational data and inject them into another operational database. The CDC is only applied over the fraction of the data that will be processed by the other operational database. The workload is not performed over the operational database due to technical or financial reasons. The technical reason is that the operational database cannot handle the full workload and some processes need to be offloaded to another database. The financial reason is that the operational database can handle the workload but the price is too high. The latter typically happens with the mainframe. We name this architecture **operational database offloading**.

A very typical and important workload that lies in having to ingest high volumes of detailed data and compute recurrent aggregate analytical queries over this detailed data. This workload has been addressed with more specific architectures. One of such architectures uses an operational database for ingesting the detailed data, and uses another operational database to store aggregated views of the data. These aggregated views are generated periodically by means of an ETL process that traverses the data from the previous period in the detail operational database, computes the aggregations and store them in the aggregate operational database. The recurrent queries are processed over the aggregation database. Since the database contains already the pre-computed aggregates the queries are light enough to be computed at an operational database. We call this architecture **detail-aggregate view splitting**. One of the main shortcomings of this architecture is the fact that the aggregate queries have an obsolete view of the data since they miss the data from the last period. Typical period lengths go from 15 minutes to hours or a full day. Some times this architecture is solved.

The kind of operational databases typically used for the above architecture are SQL operational databases and since they do not scale, they require to use an additional architectural pattern that we call **database sharding**. Sharding lies in overcoming the lack of scalability or linear scalability of an operational database by storing fractions of the data on different database independent servers. Thus, each database server handles a workload small enough, and by aggregating the power of many different database manager instances the system can scale. The main shortcomings of this architecture lie in that now queries cannot be performed over the logical database, since each database manager instance only knows about the fraction of data it is storing and cannot query any other data. Another major shortcoming lies in that there are no transactions across database instances meaning that is stored data across different instances are related, they don’t have consistency guarantees neither in the advent of concurrent reads and writes or in the advent of failures.

Other systems tackle the previous problem of recurrent aggregate queries by computing the aggregates on the application side using the memory. So basically, this in-memory aggregates are computed and being maintained as time progresses. The recurrent aggregation queries are solved by reading this in-memory aggregations, while access to the detail data are solved by reading from the operational database, many times using sharding. We name this architectural pattern **in-memory application aggregations**.

Another approach to deal with recurrent aggregate queries at scale lies in what we call **federated aggregations**. The architectural pattern lies in using sharding to store fractions of the detail data and then use a federator at application level that basically queries the individual sharded database managers getting the resultsets of the individual aggregate queries and then, aggregate them manually to compute the aggregate query over the logical database. This architectural pattern is applied frequently for monitoring systems and it is called in that context Monitor of Monitors (MoM).

In INFINITECH we envision a holistic solution to the issue of data pipelining that works with all kinds of storage, handling efficiently aggregates, and addresses the need for temporal storages to deal with snapshot databases. This holistic solution aims at minimizing the number of storage systems needed to develop an analytical pipeline and addressing all of the above identified architectural patterns for data pipelining. It also aims at unifying the data pipelines combining data streaming and data at rest.

In what follows we provide the list of targeted architectural patterns for data pipelining and how they will be automated and solved with INFINITECH's innovations in the data management layer, which are incorporated into the INFINISTORE data store

1. **Lambda architecture.**

   In INFINITECH the lambda architecture is totally trivialized by removing the at least three data management technologies and three different code bases with ad hoc code for each of the queries and just having a single database manager with declarative queries in SQL. The lambda architecture is simply substituted by the INFINISTORE, which relies on the LeanXcale database. LeanXcale scales out linearly its operational storage solving one of the key shortcomings of operational databases that motivate the lambda architecture. The second obstacle from operational databases was its inefficiency in ingesting data that makes them too expensive even for data ingestions they can manage. As the database grows, the cache is rendered ineffective and each insert requires to read a leaf node that requires first to evict a node from the cache and write to disk. This means that every write requires two IOs. LeanXcale solves this issue by providing the efficiency of key-value data stores in ingesting data thanks to the blending of SQL and NoSQL capabilities due to the use of a new variant of LSM trees. With this approach, updates and inserts are cached in an in-memory search tree and periodically propagated all together to the persisted B+ tree. Thanks to this approach the locality of updates and inserts on each leaf of the B+ tree is greatly increased amortizing the cost of each IO among many rows. The third issue solved by INFINISTORE that is not solved by the lambda architecture, it is the ease to query. The lambda architecture requires developing programmatically each query with three different code passes for each of the three layers. Using the INFINITECH data management layer and its INFINISTORE, queries are written in simple and widely known SQL. SQL queries are automatically optimized unlike the programmatic queries in the lambda architecture that require manual optimization across three different code basis for each of the layers. The fourth issue that is solved is the one of the cost of recurrent aggregation queries. In the lambda architecture, this issue is typically solved in the speed layer using data streaming. In INFINISTORE, with the development and adaption of the online aggregates, we enable real-time aggregation without the problems of operational databases and providing a low cost solution with low response time.

2. **Current-Historical Data Splitting.**

   In this approach, data is split between an operational database and a data warehouse or a data lake. The current data is kept on the operational database and historic data in the data warehouse or data lake. However, queries across all the data are not supported with this architectural pattern. In INFINITECH a new pattern will be used to solve this problem named **Real-Time Data warehousing.** This pattern will be solved by a new innovation that will be introduced in LeanXcale, namely, the ability to split analytical queries over LeanXcale and an external data warehouse. Basically, it will copy older fragments of data into the data warehouse periodically. LeanXcale will keep the recent data and some of the more recent historical data. The data warehouse will keep only historical data. Queries over recent data will be solved by LeanXcale, and queries over historical data will be solved by the data warehouse. Queries across both kinds of data will be solved in the following way. If they do not contain joins, basically the query will be executed with a predicate over time on both databases guaranteeing a split without overlapping and without missing any data item and the union of both results will be returned as the result of the query. If the query contains joins then it will be split into four subqueries. One subquery doing the joins across recent data that will be pushed down to LeanXcale. One subquery doing joins across historical data that will be pushed down to the data warehouse. And a third subquery doing joins across recent and historical data that will be solved at LeanXcale that will used the data warehouse as external data source for reading the data. In this way, the bulk of the historical data query is performed by the data warehouse, while the rest of the query is performed by LeanXcale. This approach enables to deliver real-time queries over both recent and historical data giving a 360 degree view of the data.

3. **Data warehouse offloading.**

In data warehouse offloading due to the saturation of the data warehouse data marts are used using other database managers and making a more complex architecture that requires multiple ETLs and copies of the data. Within INFINITECH this issue can be solved in two ways: One way is by using operational database offloading to INFINISTORE with the dataset of the data mart. The advantage of this approach with respect to data warehouse offloading is that the data mart contains data that is real-time, instead of obsolete data copied via a periodic ETL. The second way is to use database snapshotting taking advantage of the fast speed and high efficiency of loading of LeanXcale. In this way, a data mart can be created periodically with the same or higher freshness than a data mart would have. The advantage is that the copy would come directly from the operational database instead of coming from the data warehouse thus resulting in fresher data.

4. **Database snapshotting.**

In INFINITECH database snapshotting can be actually be avoided by using its data management layer as the operational database. This can be done thanks to the linear scalability of the INFINISTORE that does not require offloading part of the workload to other databases. However, in many cases, organizations are not ready to migrate their operational database because of the large amount of code relying on specific features of the underlying database. This is the case with mainframes with large Cobol programs and batch programs in JCL. In that case, INFINITECH by relying on LeanXcale can provide a more effective snapshotting or even able to substitute snapshotting by operational database offloading that provides full real-time data. In the case of snapshotting, thanks to the efficiency and speed of data ingestion of LeanXcale, snapshotting can be performed daily instead of weekly since load processes that takes days are reduced to minutes. But snapshotting can be substituted by operational database offloading thanks to the scalability and speed of ingestion of LeanXcale. The main benefit is that data freshness changes from weekly to real-time. This speed in ingestion is achieved thanks to LeanXcale capability of ingesting and querying data with the same efficiency independently of the dataset size. This is achieved by means of bidimensional partitioning. The bidimensional partitioning exploits the timestamp in the key of historical data to partition tables on a second dimension. Tables in LeanXcale are partitioned horizontally through the primary key. But then, they are automatically split on the time dimension (or an auto-increment key, whatever is available) to guarantee that the table partition fits in memory and thus, the load becomes CPU bound and thus, very fast. Traditional SQL databases get slower as data grows due to the B+ tree used to store data becomes bigger in both terms of number of levels and number of nodes. LeanXcale thanks to bi-dimensional partitioning keeps the time to ingest data constant. Queries are also speeded up thanks to intra-operator parallelization of all algebraic operators below joins.

5. **Operational database offloading.**

One of the main limitations of the operational database offloading is the fraction of data offloaded to a single database. Typically, this approach is performed with mainframes that can process very high workloads that soon overload other operational databases with much more limited capacity and incapable of scaling. Again by relying on LeanXcale, INFINITECH will overcome these limitations. LeanXcale can even support to full set of changes performed over the mainframe thanks to its scalability so it does not set any limitation on the dataset size and rate of data updates/inserts over this dataset.

6. Detail-aggregate view splitting.

In INFINITECH this pattern is totally removed because it is not needed anymore. By taking advantage of its declarative real time analytical framework and the so called *online aggregates* developed under the scope of T5.3, aggregate tables are built incrementally as base data is inserted. This implies to increase the cost of ingestion, but since LeanXcale is more than one order of magnitude more efficient than the market leader, it means that it can still ingest the data more efficiently despite the online aggregation, but then, recurrent aggregation analytical queries become costless since they only have to read a single row or a bunch of rows to provide the answer thanks to the fact that each aggregation has been already computed incrementally.

7. Database sharding.

Database sharding is not needed in INFINITECH thanks to the linear scalability of its INFINISTORE. Thus, what before required programmatically splitting the data ingestion and data queries across independent database instances, now, it is not needed anymore. LeanXcale is able to scale out linearly to hundreds of nodes.

8. In-memory application aggregations.

In INFINITECH in-memory application aggregations are not needed anymore removing all the problems around them like the loss of data in the advent of failures and what is more all the development and maintenance cost of the code required to perform the in-memory aggregations. Not only that in-memory aggregations work as far they can be computed in a single node, when multiple nodes are required they become extremely complex and in most cases out of reach of technical teams. In INFINITECH the online aggregates from LeanXcale will be leveraged to compute the aggregations for recurrent aggregation queries. LeanXcale keeps internally the relationship between tables (called parent tables) and aggregate tables built from the inserts in these tables (called child aggregate tables). When aggregation queries are issued, the query optimizer has been enriched with new rules to automatically detect which aggregations on the parent table can be accelerated by using the aggregations in the child aggregate table. This results in transparent improvement of all aggregations in the parent table by simply declaring a child aggregate table (obviously of the ones that can exploit the child table aggregates). More information can be found at the relevant D5.4 and D5.5 deliverables (“Framework for Declarative and Configurable Analytics”)/

9. Federated aggregations.

Federated aggregations share the motivation of in-memory aggregations but basically enable them to extend to a multiple set of nodes. As with in-memory application aggregations, INFINITECH fully solve the problem in a trial way by relying on its online aggregates.

10. Streaming data and data at rest.

Several applications require to combine streaming data with data at rest. In INFINITECH these applications will be addressed by using different mechanisms. When streaming data needs to be correlated with persistent data it will be attained by means of the integration of INFINISTORE with Flink. When streaming data needs to produce a persistent output, it will also be addressed by means of the Flink and INFINISTORE integration. However, sometimes this integration results complex because it implies writing queries in two different subsystems and it is complex their integration. For this reason, we will develop an integration of SQL with an SQL-like query language for streaming data in the form of a query language that integrates both the access to data at rest and the access to streaming data. By using a unified language, it becomes trivial the use of a column or set of columns from a streaming tuple in a query performed over the persistent storage and similarly, to integrate the output of an SQL query into the output stream of a data streaming operator that correlates streaming data with persistent data.

11. **NoSQL and SQL data.**

As previously discussed organizations have a myriad of different kinds of database managers that include both SQL and NoSQL databases. The main issue is that data stored on each family of data stores belongs to a single logical database of the organization and this split is artificial due to the technical limitations of different database technologies that prevent from using a single database for all kinds of data. In INFINITECH polyglot support will be provided to solve the data pipelines across different families of databases. Polyglot data support will enable to query from a common endpoint to SQL data stored in LeanXcale or other SQL databases and data stored in key-value data stores, wide-column data stores, document-oriented data stores, and graph databases.

Finally, INFINITECH will also integrate the different tools require to support all the data pipelines including Change Data Capture (CDC) systems and ETL tools to provide a holistic solution to the automation of data pipelining.

5 Parallized Data Stream Processing using Apache Flink and Kubernetes

In this section we will discuss the second main innovation that will be developed as part of T3.4, namely dynamically scalable stream query processing in distributed query processing environments. More precicely, we first state the motivatiton and business needs that our work will solve, the problem definition and finally, we give details of the overall design of the system using Apache Flink and Kubernetes. With our design, three key innovations are required to be achieved and we give a more detail description of the overall system covering all three of them.

5.1 Motivation

In the finance domain, there are a wide range of use-cases that require real-time processing of data streams to add value. For instance, when performing financial trading for currencies or stocks, it is critical to be able to monitor price fluctuations in real-time to identify buy/sell opportunities. Moreover, as the application of alternative information streams such as news articles and social media become more popular, large volumes of specialist compute resources are needed to enable real-time language analytics. However, a key feature of financial data is that the rate at which it arrives at is not constant. Over the course of each day the number of financial trades can fluctuate wildly, and moreover can experience bursts of activity when the market becomes aware of some new information. Indeed, in today’s trading environments, the severity of trade burstiness is exacerbated by automatic trading algorithms that use other buy/sell transactions as triggers for their own trades. As a result of these factors, to enable consistant processing of financial data streams with low latencies the underlying infrastructure needs to be elastic to rapid changes in input rate/velocity.

However, elasticity is not a feature currently supported by stream processing platforms such as Apache Flink or Spark. More precicely, such platforms provide what we will refer to as cluster scalability, i.e. compute resources in the form of worker nodes can be dynamically added or removed from the cluster, allowing the total resouces available in the cluster to be altered in real-time. On the other hand, the actual stream processing pipeline(s) deployed on the cluster are static, i.e. once configured the resources allocated to them are fixed. Hence, from a practical perspective, these stream processing platforms can’t easily tackle data streams with large fluctuations in rate/velocity effectively out-of-the-box.

Instead, what we would want is a platform that provides both cluster scalability and dynamic compute pipelines, where the development of such a platform is one of the aims of T3.4.

5.2 Problem Definition

To formalize the problem being investigated, we are considering stateful multi-operator stream processing applications over bursty data streams, such as currency trading or financial analytics. To clarify the terminology:

- **Stream Processing**: Data points needing processed will arrive over time and need to be processed as quickly as possible as they are used by down-stream components (e.g. user-facing interfaces or machine learned models).
- **Stateful**: One or more operators within the pipeline accumulate state over time that forms a part of the computation performed by those operators. This state needs to be transferred to new instances of that operator during scaling.
- **Multi-Operator**: The pipelines can contain multiple data map, reduce or transformation operations with different performance characteristics.

- **Bursty Data Stream**: The number of data points that arrive on the stream can rapidly vary across different time periods (e.g. by multiple standard deviations).

5.3 System Design

To solve this challenge, the Apache Flink query processing system developed as part of T3.3 and enhanced to enable intelligent data pipelines as described in the previous section, will be further extended to enable dynamic compute pipeline scaling. The core concept for this system is as follows. Given a point in time t, an alert is fired indicating that a pipeline (A) will soon be unable to maintain low processing latencies for a given input stream due to increased load. First, a Flink cluster with more resources will be allocated from the Kubernetes cluster and a new pipeline B will be initalized on it. This pipeline will not recieve traffic until pipeline A hits its next checkpoint. When pipeline A recieves its next checkpoint the flink source for pipeline A will be disabled, allowing for pipeline A to drain. As the snapshots from each operator are written to the persistant store, pipeline B reads and uses that data to initalize its own operators. Once all operators in pipeline B are initalized then the source for pipeline B is enabled, completing the move between pipelines. At this point pipeline A and its associated resources will be freed on the kubernetes cluster.

There are three primary innovations required to achieve the above process flow, namely: 1) enabling operator state save/load on demand between replica sets of different sizes; 2) programatic scaling of Flink clusters on Kubernetes and 3) pipeline configuration transition between clusters. We summarize each on more detail below:

5.3.1 Operator State Saving and Loading

Recent versions of Apache Flink already support checkpointing of operator states via asynchromous barriers on the input stream. Under this model, a central coodinator periodically injects barriers into the data stream, where a barrier represents a point in the stream to effectively take a ‘snapshot’ of the pipeline. When an operator recieves a barrier event, it takes a snapshot of its current state and writes that to a persistant store. If an operator has multiple inputs, it waits/blocks until it recieves the barrier from all inputs. This structure assures that the checkpoint meets termination (a complete snapshot will be produced eventualy for each input barrier) and feasibility (the snapshot will only include information up-to the barrier) guarantees. This effectively solves the state saving problem, so long as checkpointing is enabled and the snapshots are being written to a secure store then we can recover the state of each operator for different points in the stream.

However, checkpointing within Flink is designed to recover from pipeline failures (e.g. due to a machine failure), not to enable transfer of processing from a low-capacity pipeline to a higher-capacity pipeline. As such, we need to implement a new operator initalization function that enables operators spawned in a new cluster to load state snapshots from equivalent operators in a different existing pipeline. The challenge here is that since different instances of an operator may have distinct state (e.g. because they are processing different subsets of the stream), the load operator needs to understand how partitioning of the stream was performed initally to correctly set state for the new operator instances that use a different partitioning. To explain with reference to Figure 3, if we are transitioning from pipeline A to pipeline B, then for operator 1 the transfer is quite simple, as we simply need to replicate the state from operator 1 in pipeline A to both copies of that operator in pipeline B. However, for operator 2, we move from having two instances (with different states) to three instances, and hence some processing on the instance states needs to be performed to assure that the three new instances in pipeline B have the needed/correct state to function over their new partition of the input stream.

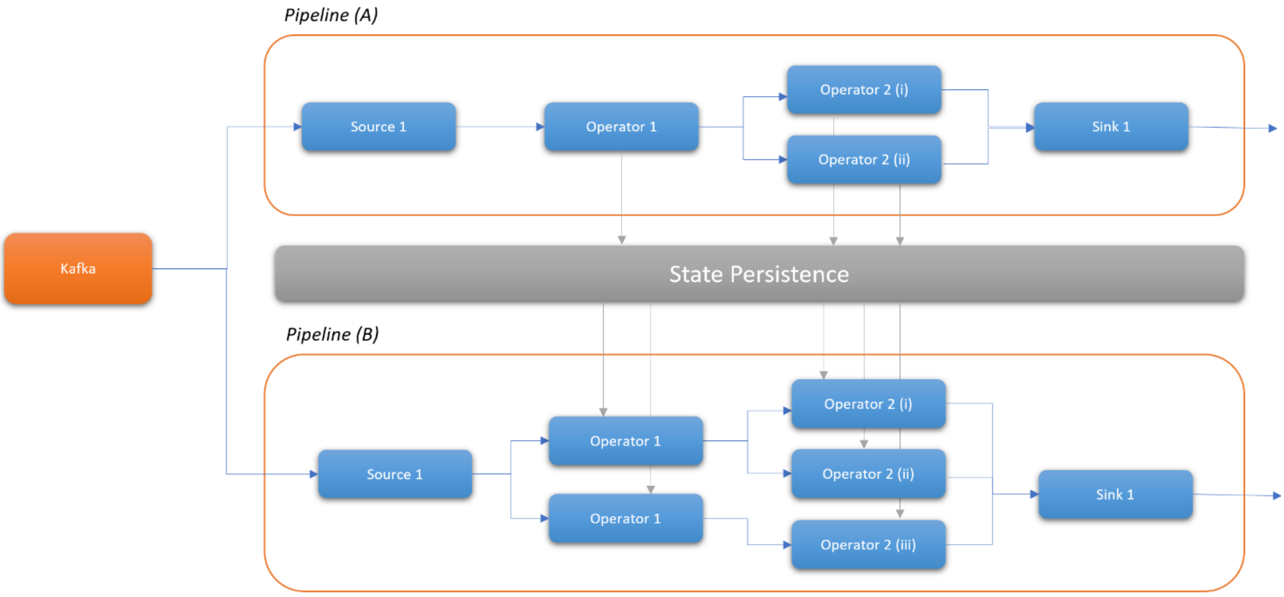

Figure 3: Pipeline Transition Diagram

5.3.2 Programmatic Scaling of Flink Clusters

The second main innovation needed to support scaling of stateful multi-operator flink applications is the programmatic scaling of the underlying Flink Cluster. Under our core design, we do not scale an existing flink cluster (although this is possible), but instead allocate a new cluster from first principles with the desired resources, followed by the deletion of the previous cluster when its no longer needed. To achieve this, a separate microservice will be built to facilitate this using the Kubernetes Operator Pattern. Kubernetes Operators are in effect software extensions to Kubernetes to enable automatic management of particular applications and their components. In this case, a kuberntes operator needs to be developed that is able to:

- Construct docker images encoding a processing pipeline
- Create and configure a new Flink Cluster with a defined total resource allocation
- Delete an existing cluster without loosing the underlying checkpoints of that pipeline

A Kubernetes Operator is itself a separate containerized service with an API that allows functions to be triggered. In this case, one function for each of the three pieces of desired functionality. The operator then communicates with the underlying Kubernetes API to operationalize the changes needed on the physical cluster infrastructure. For example, for our first function, this involves the launching of a Kubernetes Pod that executes the Docker build process to construct and then upload a container image containing our Flink compute pipeline to a docker image repository (which can be later used to produce a new Flink cluster pre-loaded with the desired compute pipeline).

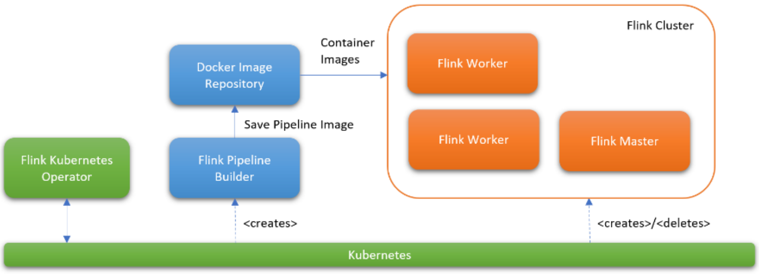

Figure 4: Flink Operator and Flink Cluster Creation

5.3.3 Pipeline Configuration Transition between Flink Clusters

The final innovation that is needed to enable scaling of stateful multi-operator Flink applications is a controller service to both trigger and then manage the overall transfer between Flink pipelines. In effect, this control service needs to provide the following functionality:

- Rule-based Predictive Pipeline Failure Identification: The ability to define a set of rules that take as input a recent set of time-series metrics exported by a compute pipeline and identify whether that pipeline needs to scale up or down. Internally, the service will regularly evaluate the different user-defined rules to see if any are violated. If so, a pipeline transition will be started.
- Pipeline Creation: If a pipeline transition is triggered, then the first action the controller service needs to perform is the creation of a new Flink cluster with appropriate resources. This is achieved through communication with the Flink Operator service defined in Section 5.3.2. Note that this operation may takes some time, and so this process needs to block until the operator reports the pipeline is in a running state.
- Data Stream Redirection: Once the new pipeline is operational, the data stream needs to be re-directed to the new processing pipeline.
- Pipeline Deletion: The final step in performing a transition between pipelines is removing the previous pipeline, via communication with the Flink Operator service defined in Section 5.3.2.

6 Conclusions and next steps

This document reported the work that has been currently done in the scope of task T3.4 "Automated Parallelization of Data Streams and Intelligent Data Pipelining", whose objective is twofold. Firstly, to provide the enablers for deploying intelligent data pipelining, thus having what we call the INFINITECH approach for intelligent data pipelines. This will provide an holistic solution that addresses all problems that currently appear in different architectural designs used in the modern landscape. The base will be the innovations brought by the data management layer of INFINITECH, which solve the problem of data ingestion in very high rates, removing the need for database offloading, along with the *online aggregates* of the declarative real-time analytical framework of INFINITECH. This removes all issues having to precalculate the results of complex analytical queries, which leads to inconsistent and obsolete results. The integration of INFINISTORE with Apache Flink, as part of the work being currently done under the scope of T3.3 "Integrated Querying of Streaming Data and Data at Rest" and the integration with tools for Change Data Capture (CDC) that will done under the scope of this task, will enable the deployment of such intelligent data pipelines.

The second objective of task T3.4 is to provide the means for enabling automated parallelization of data streams, allowing dynamically scale out individual operators that formulate a data stream in order to cope with diverse incoming workloads. Current solutions allow for static deployments of stateful multiple operators, but once they are deployed, they cannot be scaled out. Our design allows for operators to save and load their state, which allows to shutdown existing deployments and redeploy them increasing their instances, while transmitting the state of the formers to the new ones. The use of Kubernetes as the underlying container-orchestration system for automated deployments allows to programmatically scale the Apache Flink clusters and our novel operators allow to the configuration of the transmission of the state across those clusters.

At this phase of the project, the main focus was given in implementing and delivering the baseline technologies that will create the innovation and break through the current barriers of modern organizations that request real-time processing and analytics over multiple data sources (either static *at-rest* or streaming and *in-flight*). That is the HTAP provision which enables analytical query processing over the live operational data, without the need to move snapshots of a dataset to a data warehouse, the capability of the data management layer to allows for high rate data ingestion via its dual interface, its polyglot extensions that allows query processing over federated datastores, and the *online aggregates* using declarative scripting language, ensuring data consistency at the same time in terms of database transactions. As all these technologies have been incorporated into the INFINISTORE, the integration of the latter with the Apache Flink as the baseline technology for the INFINITECH streaming processing framework was the second necessity. The automation of its deployment using container-orchestration frameworks now allows for the automated parallelization of the data streams, which is the second target objective of this task. As a result, the delivery of the those fundamental pillars was the primary focus during the first phase of the project, and the first prototypes are now already available.

During this first phase of the project, an intensive analysis of the state-of-the-art of streaming processing frameworks took place that allowed us to identify the current architectural designs of the modern landscape, along with their inherit barriers. After having this analysis, we defined the vision of the outcomes of this taks that gave valuable input to the rest of the tasks related with the data management activities of INFINITECH, and more precisely, T3.1, T3.2, T3.3 and T5.3. As these tasks have been progressed and the first prototypes have been already delivered, we are now in a position to start the implementation of our holistic solution in what we call the INFINITECH approach for intelligent data pipelines. A throughout analylis on how the innovations developed so far will solve all aforementioned barriers in current architectural decisions have been provided. Moreover, the initial design on how to allow the automated parallelization of data streams have been included in this report. Our design allows to dynamically scale out individual operators of the data stream, transferring the current state of the deployed Flink cluster to the new ones, thus, removing the barrier of having to rely on static deployments that cannot cope with diverse workloads in real time.

Having the basic pillars and design in place, during the second phase of the project task T4.3 will focus on integrating the INFINITECH innovations in order to provide both the intelligent data pipeline and the automation of parallelized data streams. This is the first version of this delivery that includes the vision and design of the proposed solution. In the second version, more details will be given on the implementation, linked with specific use cases in order to highlight how our solution actually solves the current challenges of those use cases. In the third and final version, an analytical description of its use will be provided, as part of the final demonstrator and prototype.