1 Publishable summary

LeanBigData aims at addressing three open challenges in big data analytics:

1. The cost, in terms of resources, of scaling big data analytics for streaming and static data sources;
2. The lack of integration of existing big data management technologies and their high response time;
3. The insufficient end-user support leading to extremely lengthy big data analysis cycles.

Over the last years there has been a lot of progress on the scalability of big data analytics. Google, Facebook, and Amazon already process massive amounts of data. However, the techniques used for processing these large amounts of data are extremely inefficient, consuming a tremendous amount of resources resulting in a very high total cost of ownership (TCO). The amount of resources used to process data is becoming an important concern due to the fact that public cloud data centres are becoming one of the biggest consumers of energy; 2% of the electricity produced in the US is consumed nowadays by cloud data centres. World-wide data centres consume about 1.3% of the electricity produced.

Integrating different technologies over the same set of data requires a large effort, it is ad-hoc, and increases development cost for analytics. Multiple technologies are integrated mostly via an extraction-transform-load process that results in reading and writing the whole database in a periodic manner, typically daily. This widespread approach has two important problems: It affects the QoS of the production database and it is extremely costly. In some sense, although scalable, big data analytics tend to operate mostly in batch mode resulting in poor support for business processes: For instance, the Google web search engine used map-reduce jobs to process the results of crawling the web to generate a new web index from scratch. This resulted in delays of weeks from the crawling a URL until it appeared as a result in a web search. Today, Google uses a transactional index that is continuously updated and the delay has been reduced to minutes. Big data analytics still suffer from such issues in most domains.

Finally, the end-user of big data analytics is facing today long cycles of data analysis: Long cycle to discover relevant facts in data (problems, issues, alarms, etc.) that require fast reaction; Long cycle (hours or days) to get the results of large analytical queries; Long cycle to visualize the result of ad-hoc queries due to requires programmatic effort; Long cycle to interact with the visualizations until they serve the final business process.

LeanBigData is addressing these challenges by:

- Architecting and developing three **resource-efficient** Big Data management systems typically involved in Big Data processing: a novel transactional NoSQL key-value data store, a distributed complex event processing (CEP) system, and a distributed SQL query engine. The efficiency of these systems is one of the main innovations of the project. To achieve this we will remove main overheads at all levels: in the data managers (garbage collection, multi-versioning, multi-threading contention, networking, management of shared resources), in the use of underlying hardware (memory hierarchies and NUMA architectures, multi-cores, storage subsystem), in the operating system and virtualization layer, and by taking into account emerging storage technologies and trends in non-volatile memories.
- Providing **an integrated big data platform** with these three main technologies used for big data, NoSQL, SQL, and Streaming/CEP that will improve response time for unified analytics over multiple sources of data avoiding the inefficiencies and delays introduced by existing ETL-type approaches. To achieve this we will use fine-grain intra-query and intra-operator parallelism that will lead to sub-second response times for queries over static and streaming big data.
- Supporting an **end-to-end big data analytics solution** enhancing the lifecycle of data analytics by: 1) automated discovery of anomalies and root cause analysis that will provide end-users with a starting point at time 0; 2) Supporting data scientists to manipulate the result set of analytical queries in an agile way by means of a visual and interactive interface to discover insights by enabling an easy declarative manipulation of the results sets.

LeanBigData will deliver a Big Data platform that is ultra-efficient, improving today's best effort systems by at least one order of magnitude in efficiency, reducing the amount resources required to process a set of data or allowing us to process more data with the same amount of resources as today. LeanBigData will scale efficiently to 1,000s of cores. Finally, LeanBigData will demonstrate these results in a cluster with 1,000 cores in four real industrial use cases with real data, paving the way for deployment in the context of realistic business processes.