European Commission

LeanBigData

LeanBigData targets at building an ultra-scalable and ultra-efficient integrated big data platform incorporating NoSQL key-value data stores, distributed CEP systems, and distributed SQL query engines, while supporting end-to-end big data analytics by removing the main sources of delays in data analysis cycles.

AT A GLANCE

Project title:
Ultra-Scalable, Ultra-Efficient Integrated and Visual Big Data Analytics

Project coordinator:
Marta Patiño-Martinez, Universidad Politecnica de Madrid, SPAIN

Technical coordinator:
Ricardo Jimenez-Peris, Universidad Politecnica de Madrid, SPAIN

Partners:
Universidad Politecnica de Madrid (ES), INTEL (IE), Computer Associates (ES), Foundation for Research and Technology - Hellas (GR), Institute of Engineering Systems and Computers (PT), SyncLab (IT), Atos Spain S.A. (ES), Institute of Communication and Computer Systems (GR), Portugal Telecom (PT)

Duration:
36 months

Total cost:
€6,15 M

Programme:
ICT-2013.4.2 Scalable data analytics

Further information:
http://www.leanbigdata.eu

Why LeanBigData?

While over the last years there has been a lot of progress on the scalability of big data analytics, the processing techniques are extremely inefficient, *consuming a tremendous amount of resources*, and thus resulting in a very high total cost of ownership (TCO). Besides the cost, the 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 (world-wide data centres consume about 1.3% of the electricity produced).

What is more, *integration of different data management technologies* requires a large effort, while it is an ad-hoc process that increases development cost for analytics. These technologies are usually integrated via an extraction-transform-load (ETL), which however 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.

Finally, the *end-user of big data analytics* is facing today long cycles across the data analysis lifecycle: from discovering relevant facts (such as issues or alarms), to obtaining the results of large analytical queries, visualizing the result of ad-hoc queries, and interaction with the visualizations.

What makes LeanBigData unique?

LeanBigData will "***do it faster with less resources***".

LeanBigData will initially focus on the core underlying data management technologies, by architecting and developing **three resource-efficient big data management systems**: a novel transactional NoSQL key-value data store, a distributed complex event processing (CEP) system, and a distributed SQL query engine. **Ultra-efficiency** will be achieved through enhanced data managers, hardware-related techniques, approaches oriented on the operating system and virtualization layers, and technologies focusing on storage and non-volatile memories.

Furthermore, LeanBigData will provide 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 this end, LeanBigData 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.

Accelerating the data analysis cycles will also be achieved through LeanBigData approaches, by supporting an **end-to-end big data analytics solution** that will remove the four main sources of by using: 1) automated discovery of anomalies and root cause analysis that will provide end-users with a starting point at time 0; 2) incremental visualization of results of long analytical queries to allow discarding inappropriate queries without waiting until the results are delivered hours or days later; 3) drag-and-drop declarative composition of visualizations; and 4) efficient manipulation of visualizations and query results through hand gestures over 3D/holographic views

Value proposition

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** and will enable **unified processing in real-time** of millions of streaming events per second and queries over trillions of records with subsecond results.

Demonstrators

The LeanBigData outcomes will be validated through four real industrial use application scenarios: ***Cloud Data Centres Monitoring***, to monitor and correlate application performance, hardware and data centers utilization and predict failures; ***Targeted Advertisement*** to maximize impact and return through real-time queries with respect to advertisements; ***Alignment of Financial Direct Debit Transactions*** to detect direct debit frauds in a timely and efficient way; and ***Social Network Analytics*** to analyse at real-time and visualize social media graphs.