# The Sigma Data Processing Architecture: Leveraging Future Data for Extreme-Scale Data Analytics to Enable High-Precision Decisions

Gabriel Antoniu[1], Alexandru Costan[1], Maria S. Pérez[2], Nenad Stojanovic[3]

[1]Univ Rennes, Inria, CNRS, IRISA

[2]Universidad Politécnica de Madrid

[3]Nissatech

6 November 2018

This white paper introduces several key principles based on which HPC-Big Data convergence can be achieved: 1) use future (simulated) data to substantially enrich knowledge obtained based on historical (past) data; 2) enable high-precision analytics thanks to hybrid modeling combining simulation and data-driven models; 3) enable unified data processing thanks to a data processing framework able to relevantly leverage and combine stream processing and batch processing in situ and in transit.

Due to an ever-growing digitalization of the everyday life, massive amounts of data start to be accumulated, providing larger and large volumes of historical data (**past data**) on more and more monitored systems. At the same time, an up-to-date vision of the actual status of these systems is offered by the increasing number of sources of real-time data (**present data**). Today's data analytics systems correlate these two types of data (past and present) to predict the future evolution of the systems to enable decision making. However, the relevance of such decisions is limited by the knowledge already accumulated in the past.

At the same time, companies and organizations do not only want to learn from the past, but also aim to get a precise understanding on what might happen next in any circumstances, in particular how to react to unknown events. A timely and relevant example is the connected vehicle (e.g., vessel or car) with autonomy facilities: on one side, computational simulation models [Chinesta2013] are used to simulate the vehicle's behavior in various hypothetical conditions in order to improve its design; on the other side, data-driven analytics models [Ibanez2017] are used to monitor and control the system in real time during its operation, to support vehicle motion through advanced sensing. Combining knowledge from both models appears as a high-potential opportunity to substantially improve the vehicle's performance and maintenance process.

More generally, we witness a huge expansion of complex systems and processes where the knowledge acquired based on **past data** can substantially be extended with what could be called **future data** generated by simulations of the system behavior under various hypothetical conditions that have not been met in the past. This can provide a richer tool for much deeper interpretation of measured real-time data, enabling much more relevant decision making, beyond what is currently enabled by history-based prediction methods.

**Challenges.** Enabling the use of the future data jointly with past and present data leads to several challenges:
The overall challenge is to define and validate the most appropriate methodology that supports the combination of the two modelling paradigms in a way that exploits the potential for synergies. This translates into:

- A **challenge regarding the data analytics model**: combine computational and data-driven analytics models through hybrid modeling.
- A **challenge regarding the data processing architecture**: efficiently integrate simulations and data analytics through a unified data processing architecture combining traditional Big Data processing (batch- and stream-based) with HPC techniques for data processing (in situ, in transit) to support hybrid analytics models.
- **A challenge regarding continuous model improvement**: simultaneously refine both the data-driven model and the computational model through continuous learning, with the goal of optimizing the real-world system behavior.

**Future Data: from Big Data to Extreme Data.** Combining computation-driven and data-driven analytics can reach full potential only if the two types of data analytics efficiently leverage each other to detect the best opportunities to improve the system operation, but also to react in an optimal way to critical situations. This leads to high challenges related to the extreme scale of data management in terms of both volume and velocity.

- First, the number of hypothetical scenarios, the possibility to simulate them with a virtually unlimited combination of parameters and the possibility to run joint data analysis correlating such hypothetical data with past measured data as well as with real-time data coming from the real system may produce immense amounts of data to process (**extreme volume**).
- At the same time, performing such a complex analysis on a virtually immense number of scenarios in order to find the most efficient way to react in real time to some critical situations that require very fast decision poses a challenge in terms of **extreme velocity** for data processing, which can require extreme-scale computations performed on extreme-scale HPC infrastructures.

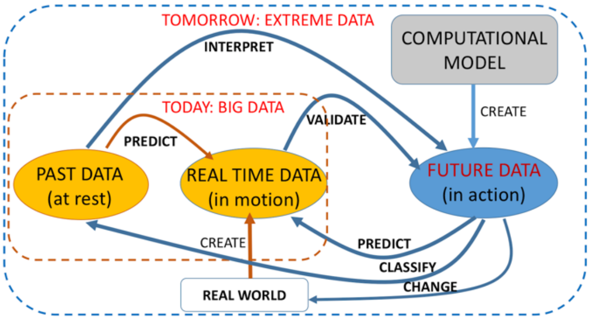

*Figure 1. From Big Data to Extreme Data*

**The need for Hybrid Analytics Models.** So far, data-driven analytics and simulation-based modelling have been developed and used separately. They involve quite different cultures and types of expertise: numerical models typically simulated in HPC environments on one side; Big Data analytics models and technologies on clouds on the other side. The two types of modelling provide complementary views:

- Data-driven analytics enables understanding/learning the real system's behavior and its rationale from past data, leveraging machine learning/deep learning techniques;
- Simulation based on complex computational models can provide data to forecast the system's performance in various possible situations (including some that have never occurred yet).

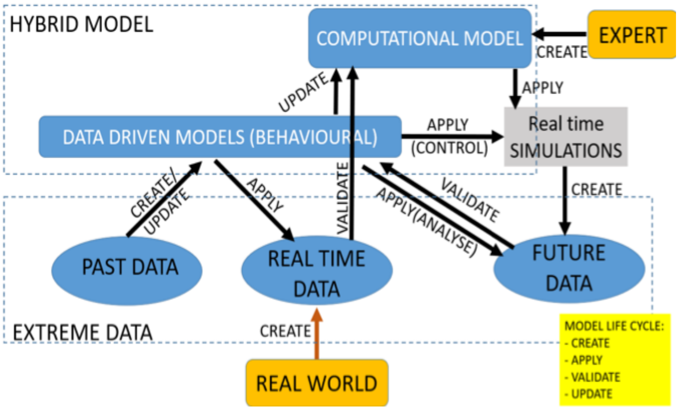

*Figure 2. Hybrid Analytics Models*

By combining information from both models, hybrid analytics models can be built: they can substantially augment knowledge for understanding and predicting the system's behavior: the knowledge based on potential future behavior will substantially enrich the knowledge acquired based on past behavior.

**Continuous model improvement.** The quality of the simulation model impacts the quality of the future data they generate and, thereby, the accuracy of the predictions of future events. To improve the simulation model, several techniques can be used. First, acquired data from the real system can be used for calibration. Second, the model can also be corrected based on experience with the system operation, by exploiting a behavioral model built based on deviations between predictions and measurements (hybrid model). Finally, in order to avoid jeopardizing real-time feedback, this process should consider only valuable, high-impact data that increases the knowledge encapsulated into the model correction. The challenge is to succeed in efficiently leveraging these relevant data as a key approach to enrich the physics-based simulation model, allowing it to produce better (faster and more accurate) future data. At the same time, the data-driven model is subject to continuous improvement, by assimilating simulation data.

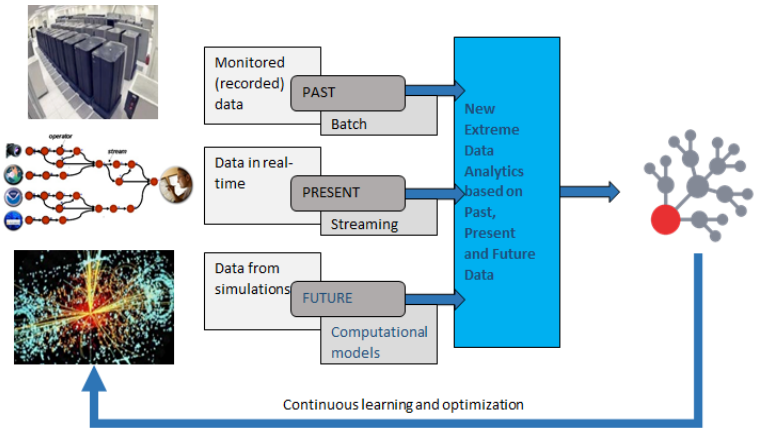
operator
stream
Monitored (recorded) data
PAST
Batch
Data in real-time
PRESENT
Streaming
Data from simulations
FUTURE
Computational models
New Extreme Data Analytics based on Past, Present and Future Data
Continuous learning and optimization

*Figure 3. Using learning to continuously update and refine the models.*

**Enabling Hybrid Analytics: the Sigma Architecture for Data Processing.** Traditional *data-driven analytics* relies on *Big Data processing* techniques, consisting of *batch processing* and *real-time (stream) processing*, potentially combined in a so-called *Lambda architecture*. This architecture attempts to balance latency, throughput, and fault-tolerance by using batch processing to provide comprehensive and accurate views of batch data, while simultaneously using real-time stream processing to provide views of online data.

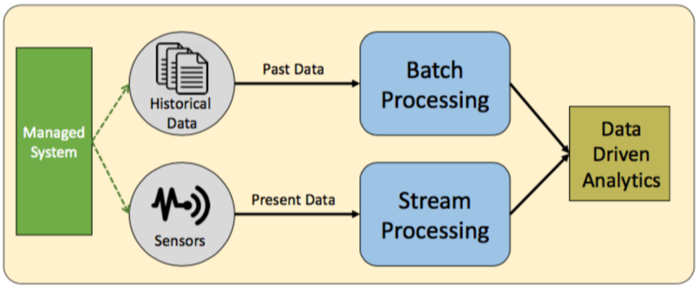

*Figure 4. The Lambda data processing architecture.*

On the other side, *simulation-driven analytics* is based on computational (usually physics-based) simulations of complex phenomena, which often leverage HPC infrastructures. The need to get fast and relevant insights from massive amounts of data generated by extreme-scale simulations led to the emergence of *in situ* and *in transit* processing approaches [Bennet2012]: they allow data to be visualized and processed interactively in real-time as data are produced, while the simulation is running.

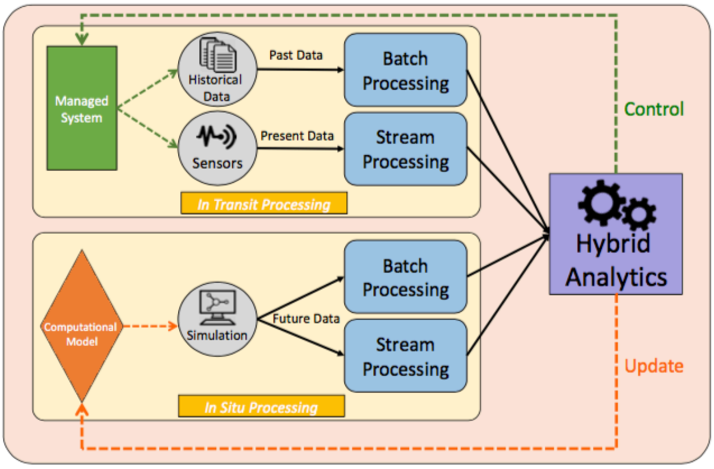

*Figure 5. The Sigma data processing architecture.*

To support hybrid analytics and continuous model improvement, we propose to combine the above data processing techniques in what we will call the **Sigma architecture**, a HPC-inspired extension of the Lambda architecture for Big Data processing. Its instantiation in specific application settings depends of course of the specific application requirements and of the constraints that may be induced by the underlying infrastructure. Its main conceptual strength consists in the ability to leverage in a unified, consistent framework, data processing techniques that became reference in HPC in the Big Data communities respectively, without however being combined so far for joint usage in converged environments.

## Conclusion

We believe the principles sketched out above put in place a sound base for making a step further towards the convergence of the HPC and Big Data Analytics areas. The Sigma architecture can be established as a new paradigm enabling convergence at data processing level. Based on it, hybrid modeling can become a reference paradigm for data analytics. We see these two paradigms as key enablers of extreme data analytics for emerging application scenarios that will efficiently leverage converged HPC-Big Data infrastructures to enable unprecedentedly high decision making in complex environments.

## References

[Chinesta2013] F. Chinesta, A. Leygue, F. Bordeu, J.V. Aguado, E. Cueto, D. Gonzalez, I. Alfaro, A. Ammar et A. Huerta. Parametric PGD based computational vademecum for efficient design, optimization and control. Archives of Computational Methods in Engineering, 20/1, 31-59, 2013.

[Ibanez2017] R. Ibanez, E. Abisset-Chavanne, J.V. Aguado, D. Gonzalez, E. Cueto, F. Chinesta. A Manifold-Based Methodological Approach to Data-Driven Computational Elasticity and Inelasticity. Archives of Computational Methods in Engineering, 2017.

[Bennet2012] J.C. Bennet, H. Abbasi, P.-T. Bremer, R. Grout et al. Combining in-situ and in-transit processing to enable extreme-scale scientific analysis. In Proc. ACM SC'12, Salt Lake City, Nov. 2012.

[Carbone2015] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, K. Tzoumas, Apache Flink: Stream and batch processing in a single engine, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 36 (4).

[Marcu2017] O. Marcu, A. Costan, G. Antoniu, M.S. Pérez, R. Tudoran, S. Bortoli and B. Nicolae, Towards a Unified Storage and Ingestion Architecture for Stream Processing. IEEE International Conference on Big Data (Big Data), 2402–2407; 2017. doi: 10.1109/BigData.2017.8258196