A Data Lake to Accelerate Real-time Analytics

Using HPE Elastic Platform for Analytics (EPA) and the MapR Data Platform

With stream processing growing at a CAGR of 32% annually, it's the fastest growing software category for Big Data Analytics. Legacy infrastructure used to support Hadoop data lakes will be inadequate for real-time data lake enablement.

HPE Elastic Platform for Analytics (EPA) platform provides a comprehensive data pipeline for Edge-to-Core-to-Data Lake infrastructure for Real-time Analytics. Some critical features offered by this platform include:

- **Modular reference architectures** to enable the independent scaling of compute and storage for Edge Analytics, Streaming Analytics, Batch Analytics, AI, and Data Lake.

- **Real-time data ingest** with edge connectors and primitive data preprocessing for Edge Analytics.

- **GPU infrastructure modules for Machine learning, Deep learning, and Neural Net** capabilities to build Edge ML, Core ML, and Data Lake ML.

HPE EPA platform is an infrastructure optimized for the **MapR Data Platform** and provides end-to-end capabilities for a robust data pipeline. Key features of the MapR Data Platform include:

- **MapR Event Store for Apache Kafka:** A Pub/sub engine for real time high speed ingestion of data

- **MapR-XD:** An enterprise-grade scalable distributed file and object store with storage volume containers

- **MapR Database:** A scalable NoSQL data storage enabling high performance interactive ETL with Apache Drill and Hive

Faster insights to drive new business opportunities

Enterprises can no longer wait weeks or even days to generate analytics and to uncover new business insights. To compete in the current market, enterprises must continuously acquire data on their customers, suppliers, partners, and competitors in their market.

The digital transformation is now delivering systems, smart devices out on the edge and new applications to capture and analyze massive datasets, all in real time. This enables businesses to leverage data from traditional data warehouses as well as new data streams coming in from the edge.

This trend is not unique to a few industries, but goes across a wide swath of industry use cases, including Financial Services, Healthcare, Manufacturing, Retail, Telecom, to name a few. The driving objective is to exploit new ingest streams and to make real-time decisions on this data, achieving new operational efficiencies, identifying new revenue streams and ultimately, improving customer satisfaction. Some of the key challenges customers face in attempting to make this transition would include:

- Lack of comprehensive infrastructure solution purpose-built from Edge-to-Core-to-Data Lake for Big Data Analytics Solutions.

- Slow performing legacy Hadoop solutions leading to infrastructure cluster sprawl.

- Lack of low latency infrastructure solution for Edge Analytics integrating to high performance requirements of streaming analytics at the core.

- Lack of a comprehensive capability to build and deploy advanced analytic machine learning models across Edge-to-Core-to-Data Lake.

- Lack of an automated, policy-based mechanism to move hot, warm, and cold data to appropriate storage tiers based on cost, performance, and capacity trade-offs.

Historically the data lake has been a prominent unified platform used to store enterprise data, including raw copies of source system data and transformed data used for tasks such as reporting, visualization, and batch analytics.

More recently, the industry requires integrating these data lakes with real-time data sources such as smart car sensors, financial transactions, machine log data, traffic sensors, geo-spatial data, social media, and web clickstreams. Many of these data sources trigger events and in turn create event streams. The data is generated at a very high volume and needs to be processed as soon as possible to provide real-time results with extremely low latency.

Ingest and analysis of such data demands a robust data pipeline with purpose built infrastructure. The majority of legacy infrastructure used today is inadequate in addressing this need. This would also require several distributed applications to be linked together in real time as newer larger datasets arrive.
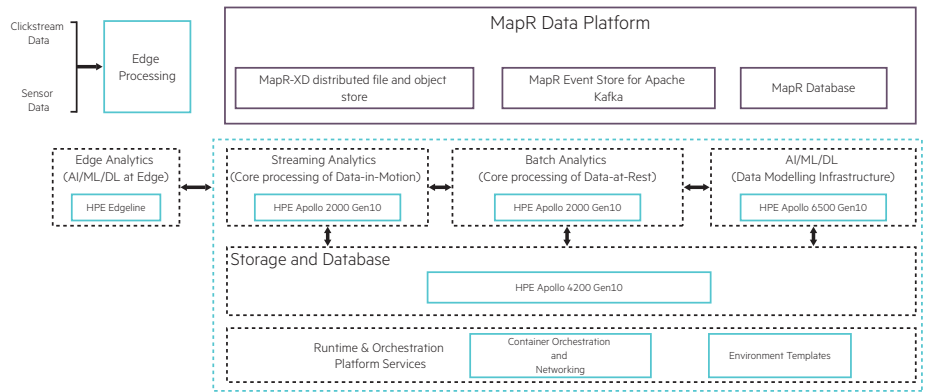
Figure 1. Real-time Analytics building blocks

HPE Elastic Platform for Analytics (EPA) platform provides a comprehensive data pipeline for Edge-to-Core-to-Data Lake infrastructure for Real-time Analytics.

MapR enables cohesive data pipeline integrated messaging, micro-batching, and data storage technology with MapR-ES, MapR-FS, and MapR-DB.

MapR-ES provides the critical functionality to enable real-time streaming, built around HPE EPA architecture.

For more information on this solution:

• **don.wilson@hpe.com**

• **info@mapr.com**

• **rchokshi@mapr.com**

Watch a MapR video on MapR Enterprise Streaming Platform

MapR Streams provides a robust data pipeline with topics that can help organize events into categories. They are logical collections of messages managed by MapR Event Store. Producers deliver data to topics and consumers subscribe to the topic to consume messages and there by analyze the messages further.

HPE Apollo systems provide purpose-built infrastructure with a high-density compute platform in a 2U chassis with HPE Apollo 2000, designed for Streaming Analytics, as well as a dense storage platform with the HPE Apollo 4200 optimized for the Data Lake and tiered storage. Finally, the HPE Apollo 6500 delivers a purpose built, highly dense GPU platform optimized for AI workloads.

Advantages of Real-time Analytics Data Pipeline with MapR and EPA

MapR enables cohesive data pipeline integrated messaging, micro-batching, and data storage technology with MapR Event Store, MapR-XD and MapR-DB. The advantages of adopting the combined MapR and HPE solution are summarized below:

• Ingesting huge volumes and variety of real-time data generated by sensors with Edge Infrastructure

• Streaming massive volume of unstructured data to scalable data pipeline with Core Infrastructure

• Persisting high velocity data (structured and unstructured) to Data Lake Storage Infrastructure

• Analyzing real-time data with machine learning models deployed at Edge, Core, and Data Lake

Learn more at
hpe.com/us/en/servers/big-data-analytics.html

Make the right purchase decision. Click here to chat with our presales specialists.

Share now

Get updates