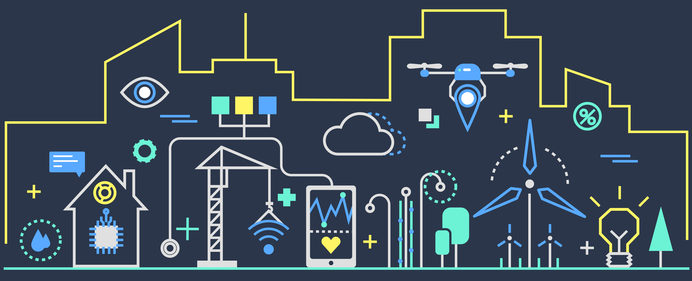
instaclustr



Powering Applications with

Apache Kafka | Apache Spark | Apache Cassandra

Today's application needs to capture, store, process, analyse, search and explore data at limitless scale and with no downtime. Successful applications need to serve millions of users and nothing less than real-time information makes sense.

Meeting these requirements needs a leading distributed technologies in each category - Apache Kafka for streaming, Apache Spark for analytics and Apache Cassandra as a data store.

kafka

Real-time message stream captures and publish data.

APACHE Spark

Data is then filtered, enriched and formatted before it is used by an application.

cassandra

Operational Database is used for application deployment and analytics.

Apache Kafka is the leading streaming and queuing technology for large-scale, always-on applications. Kafka takes streams of real-time messages, stores them reliably on a central cluster and allows those streams to be received by applications that process the messages. The stream of real-time data can come from sources like datastore, applications, Internet of things and more.

Kafka is a distributed system and can scale up and down by adding servers or instances to the cluster. Apache Kafka uses load balancing and data replication to allow failure or planned maintenance of individual nodes with no downtime.

Apache Spark is a distributed, memory-optimized system. A perfect complement to Kafka, it includes streaming library and rich set of programming interfaces to make data processing and transformation easier. Spark can detect patterns and provide actionable insight to your data. The fast and powerful open source processing engine, Apache Spark is built around speed, ease of use and sophisticated analytics.

Spark can ingest data from Kafka. Converting that into smaller data set running enrichment operations to augment data, Spark can then push that refined data set to a persistent data store.

Spark does not include a data store but almost always requires one for a complete application. Instaclustr Managed Apache Spark is colocated with Apache Cassandra, which means Spark engine is right where your operational database resides. No need for extracting, transforming and loading into a new environment. Spark also fully integrates with the key components of Cassandra and provides the resilience and scale you would need for your application.

Apache Cassandra
Real-time data streams provide the most when analysis spans both real-time and historical data. For this, the data must be persisted beyond the streaming aspects like messaging and transformation into a permanent datastore. Apache Cassandra is the open source distributed NoSQL database designed to handle large amount of data across commodity servers, providing high availability with no single point of failure and has been architected from the ground up to handle large volumes of data.