Accionlabs

Accion INNOVATION SUMMIT 2023

02-05 March 2023,
Sofitel Dubai
The Palm Jumeirah
Dubai

Accion INNOVATION SUMMIT 2023

Accionlabs

Is Spark is losing its sparkle?
How big data analytics platforms are evolving

Accion
INNOVATION
SUMMIT 2023

Accionlabs

DC

Chief Architect

An old guy who still plays around with Legacy and Modern Technologies in the process consults our customers to achieve their lofty goals of Digital Transformation.

A Hands-on Technologist.

Accion
INNOVATION
SUMMIT 2023

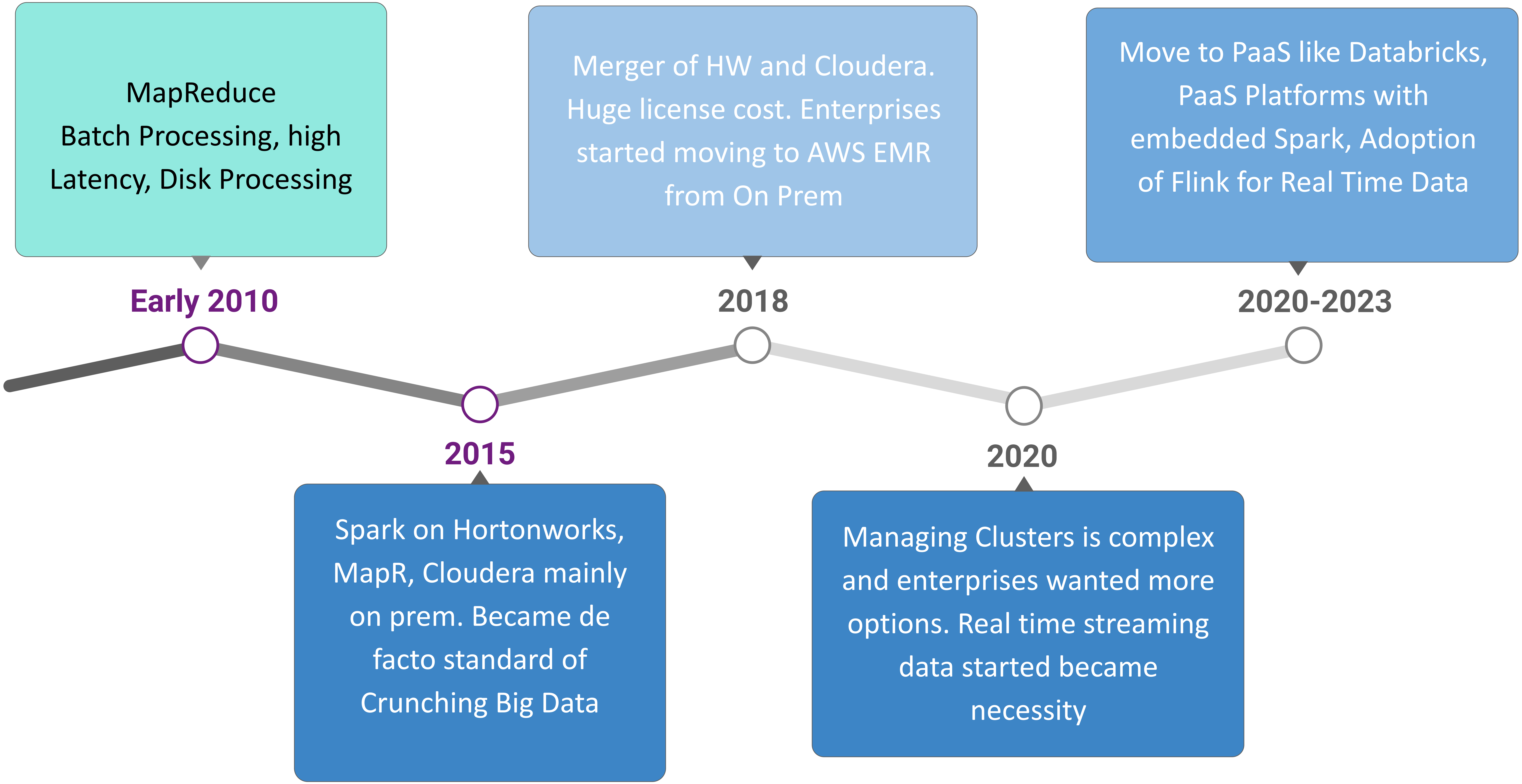
Accionlabs

Sanket Shah

Cloud Maverick

Forever learner and experimenter on multi-cloud Architecture combined with Business needs mapping. Swinging between old to new technologies through hands-on experiences.

Evolution of Spark



Early 2010

MapReduce
Batch Processing, high Latency, Disk Processing

2015

Spark on Hortonworks, MapR, Cloudera mainly on prem. Became de facto standard of Crunching Big Data

2018

Merger of HW and Cloudera. Huge license cost. Enterprises started moving to AWS EMR from On Prem

2020

Managing Clusters is complex and enterprises wanted more options. Real time streaming data started became necessity

2020-2023

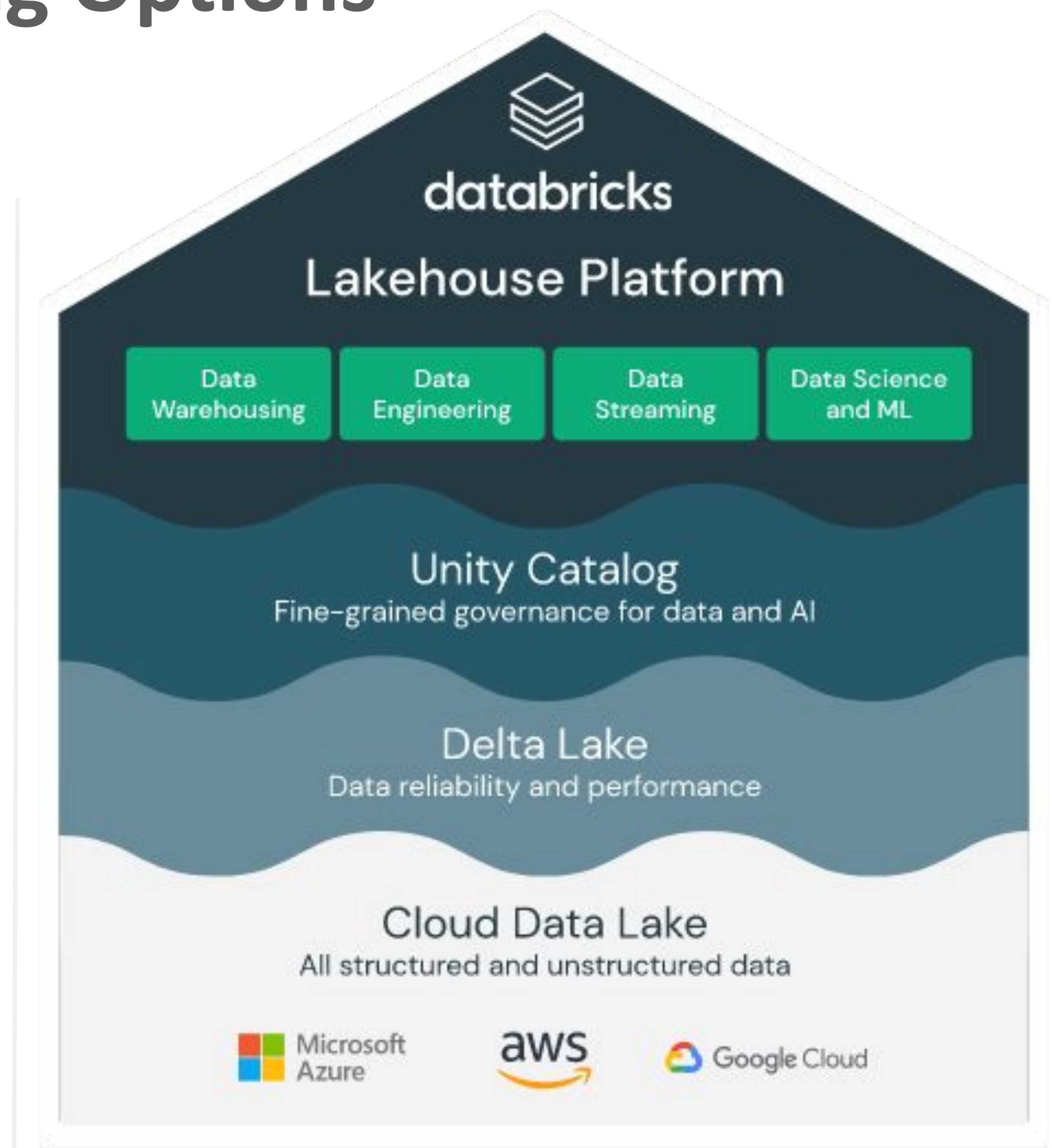
Move to PaaS like Databricks, PaaS Platforms with embedded Spark, Adoption of Flink for Real Time Data

Batch Processing Options

Apache Spark | amazon EMR

Azure HDInsight | Cloud Dataproc

Spark running on On Prem or Cloud Cluster



databricks

Lakehouse Platform

Data Warehousing | Data Engineering | Data Streaming | Data Science and ML

Unity Catalog

Fine-grained governance for data and AI

Delta Lake

Data reliability and performance

Cloud Data Lake

All structured and unstructured data

Microsoft Azure | aws | Google Cloud

Multi Cloud PaaS - Spark, MLFlow and Delta Lake

Azure Synapse Analytics

Azure PaaS - wrapper on Spark

bodo.ai

Unicorn. True HPC engine. Compiler Driven. Claims to be faster than Spark

Comparison on: Maturity | DevOps Complexity | Flexibility | Security

Batch Processing Options - Maturity

Accionlabs

APACHE Spark amazon EMR

Azure HDInsight Cloud Dataproc

databricks

Azure Synapse Analytics

bodo.ai

Spark - On-prem / Cloud

- Maintained by active OSS Community
- Cloud Providers use underlying services
- Range of manual override feature configuration options

Databricks

- Abstracts default configurations
- Provides Delta Lake, Delta Sharing, MLFlow and Redash natively
- True multi-cloud

Azure Synapse

- Features of Databricks plus:
- Microsoft Cloud specific libraries and support
- Integrates natively with Cosmos DB (NoSQL)

BODO.AI

- Bodo currently does not support native integrations to most of the databases

Batch Processing Options - DevOps Complexity

Accionlabs

APACHE Spark amazon EMR

Azure HDInsight Cloud Dataproc

databricks

Azure Synapse Analytics

bodo.ai

Spark - On-prem / Cloud

- Needs manual configuration
- CI / CD can be difficult if multiple services are used
- Secrets and configuration maintenance needs to be considered additionally

Databricks

- In-built support for Git Repository

Azure Synapse

- In-built support for Git Repository
- Secrets & configuration management can be driven externally through Azure DevOps and Azure Key Vault

BODO.AI

- Bodo Cloud managed instances does not have DevOps complexity
- On Prem complexity is similar to Spark On Prem

Batch Processing Options - Flexibility

APACHE Spark

amazon EMR

Azure HDInsight

Cloud Dataproc

Spark - On-prem / Cloud

- Completely flexible
- All options can be overridden by custom configurations and support provided by Cloud Providers

databricks

Databricks

- Semi-flexible as only few options can be changed
- Supports Cluster libraries
- Workload type based pricing

Azure Synapse Analytics

Azure Synapse

- Semi-flexible as only few options can be changed
- Supports Cluster libraries

bodo.ai

BODO.AI

- Completely flexible
- All options can be overridden by custom configurations and support provided by Cloud Providers

Batch Processing Options - Security

Spark - On-prem / Cloud

- Depends on the Cloud Platform and Team maturity
- Need to configure manually for enterprise level features, and availability may be restricted

Databricks

- Row Level Security
- Data Masking on the fly (through Fernet)

Azure Synapse

- Row Level Security
- Data Masking on the fly

BODO.AI

- Need to write custom code for row level security or Data Masking

Batch Processing Options - Verdict

- Although Enterprises are trying to move out of managed Clusters using Spark, Spark is still the Enterprise Choice for Batch Processing Albeit in a different Avatar
- Features of Spark are constantly developed and are being pulled into commercial versions and wrapper products
- Development of Spark is supported by Databricks, AWS, Microsoft and other corporations
- Depending on the business case, cloud agnostic vision and other factor, appropriate derivative of Spark can be used

Stream Processing Options

Accionlabs

APACHE Spark

amazon EMR

Azure HDInsight

Cloud Dataproc

databricks

Flink

Azure Stream Analytics

Spark based Platforms

Apache Flink

Azure Stream Analytics

<u>Comparison on:</u> Latency | Windowing | Data Processing Methodology | State Management

Stream Processing Options - Latency

APACHE Spark

amazon EMR

Azure HDInsight

Cloud Dataproc

databricks

Flink

Azure Stream Analytics

Spark based Platforms

- High (due to micro-batching)
- Increases latency for higher throughputs

Apache Flink

- Very Low
- Scales well for high throughputs

Azure Stream Analytics

- Very Low
- Scales extremely well for high throughputs

Stream Processing Options - Data Processing Methodology

Apache Spark

amazon EMR

Azure HDInsight

Cloud Dataproc

databricks

Flink

Azure Stream Analytics

Spark based Platforms

- Micro Batching
  - Batch processes on much smaller accumulations of data – typically less than a minute’s worth of data with low volumes.

Apache Flink

- Native - Streaming
  - Immediately process new records through the whole pipeline, which we need for continuous and low-latency stream processing.

Azure Stream Analytics

- Native - Streaming
  - Immediately process new records through the whole pipeline, which we need for continuous and low-latency stream processing.

Stream Processing Options - Windowing

APACHE Spark

amazon EMR

Azure HDInsight

Cloud Dataproc

databricks

Flink

Azure Stream Analytics

Spark based Platforms

- Tumbling
- Sliding

Apache Flink

- Tumbling
- Sliding
- Session
- Global
- Custom

Azure Stream Analytics

- Tumbling
- Sliding
- Session
- Snapshot

Stream Processing Options - State Management

APACHE Spark

amazon EMR

Azure HDInsight

Cloud Dataproc

databricks

Flink

Azure Stream Analytics

Spark based Platforms

- Only HDFS based file systems
- May result into out-of-memory issues as memory is shared with Executor

Apache Flink

- Memory
- File System
- RocksDB

Azure Stream Analytics

- Checkpointing
- Query Partitions

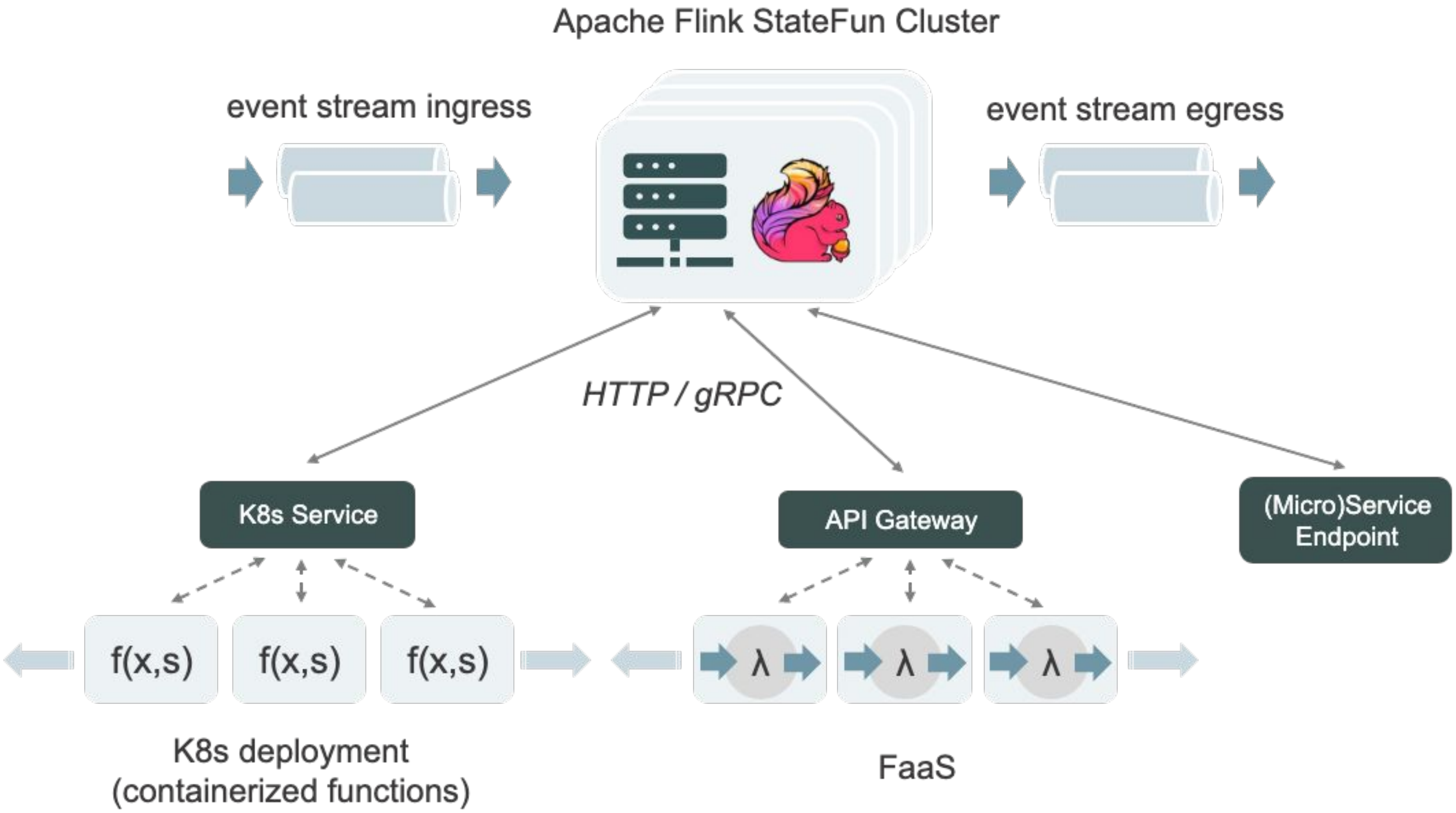
Stream Processing Options - Verdict

- Spark is still a choice when throughput is low and near real time suffices the need
- Enterprises are trying to move out of Spark in the Real Time streaming world
- Azure Stream Analytics is also being adopted at a very fast pace for following reasons:
  - Easy to Setup - can be hosted on cloud or on-premises
  - Easy to Use - SQL style support
  - Can be used with Azure Functions for CEP (Complex Events Processing)
  - Supports C# and JavaScript for extensibility
- Flink has matured over time and is becoming a CTO’s choice because of native streaming and stateful functions for following reasons:
  - Cloud Agnostic and Containerization support
  - More Flexible as developers can extend all the functions in Java or Scala

Stateful Functions: A Platform-Independent Stateful Serverless Stack

- A simple way to create efficient, scalable, and consistent applications on modern infrastructure - at small and large scale
- https://nightlies.apache.org/flink/flink-statefun-docs-stable/
- Stateful Functions
  - is an API that simplifies the building of distributed stateful applications with a runtime built for serverless architectures.
  - It brings together the benefits of stateful stream processing - the processing of large datasets with low latency and bounded resource constraints
    - along with a runtime for modeling stateful entities that supports location transparency, concurrency, scaling, and resiliency.

Stateful Functions: Architecture



Demo

Accion INNOVATION SUMMIT 2023

Accionlabs

No, Spark has NOT lost its Sparkle

Accion INNOVATION SUMMIT 2023

Accionlabs

Short Demo to show how easy is today to use a spark cluster

Accionlabs

Accion INNOVATION SUMMIT 2023

Thank you!!!

Please reach out us for discussing more at:

DC (Dwaip Chowdhury)
dc@acciolabs.com| +91 93410 19168
Sanket Shah
sanket.shah@accionlabs.com | +91 98793 56075