AWS SUMMIT

Serverless Big Data Architectures

Serverless Streaming Data Analytics

Ben Snively, AWS Sr. SA, Data & Analytics

April 19, 2017

Agenda

Cloud architecture evolution – Why serverless

Data and analytics flow

Key services overview

Design patterns

Call to action

Cloud architecture evolution

Virtualized → Managed → Serverless

Virtualized servers

Managed platforms

Serverless analytics

Serverless characteristics

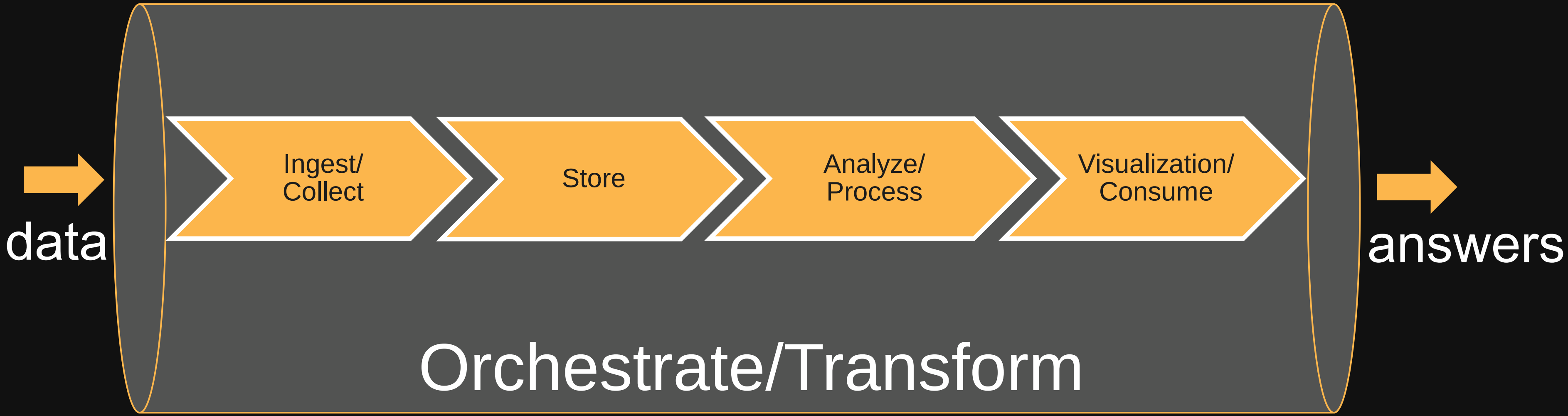
No servers to provision or manage

Scales with usage

Never pay for idle

Availability and fault tolerance built in

Data and analytics flow



data

Ingest/ Collect

Store

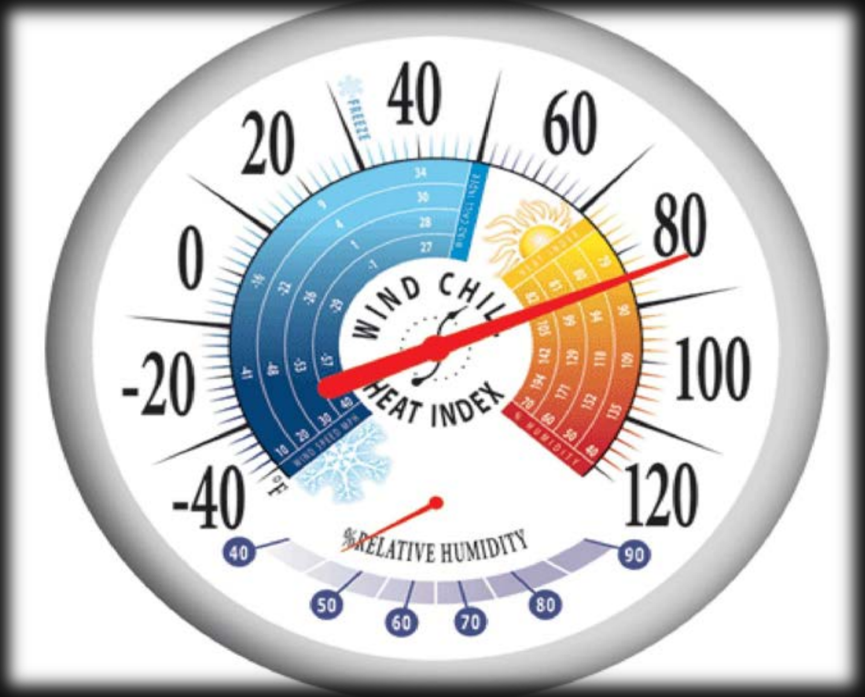
Analyze/ Process

Visualization/ Consume

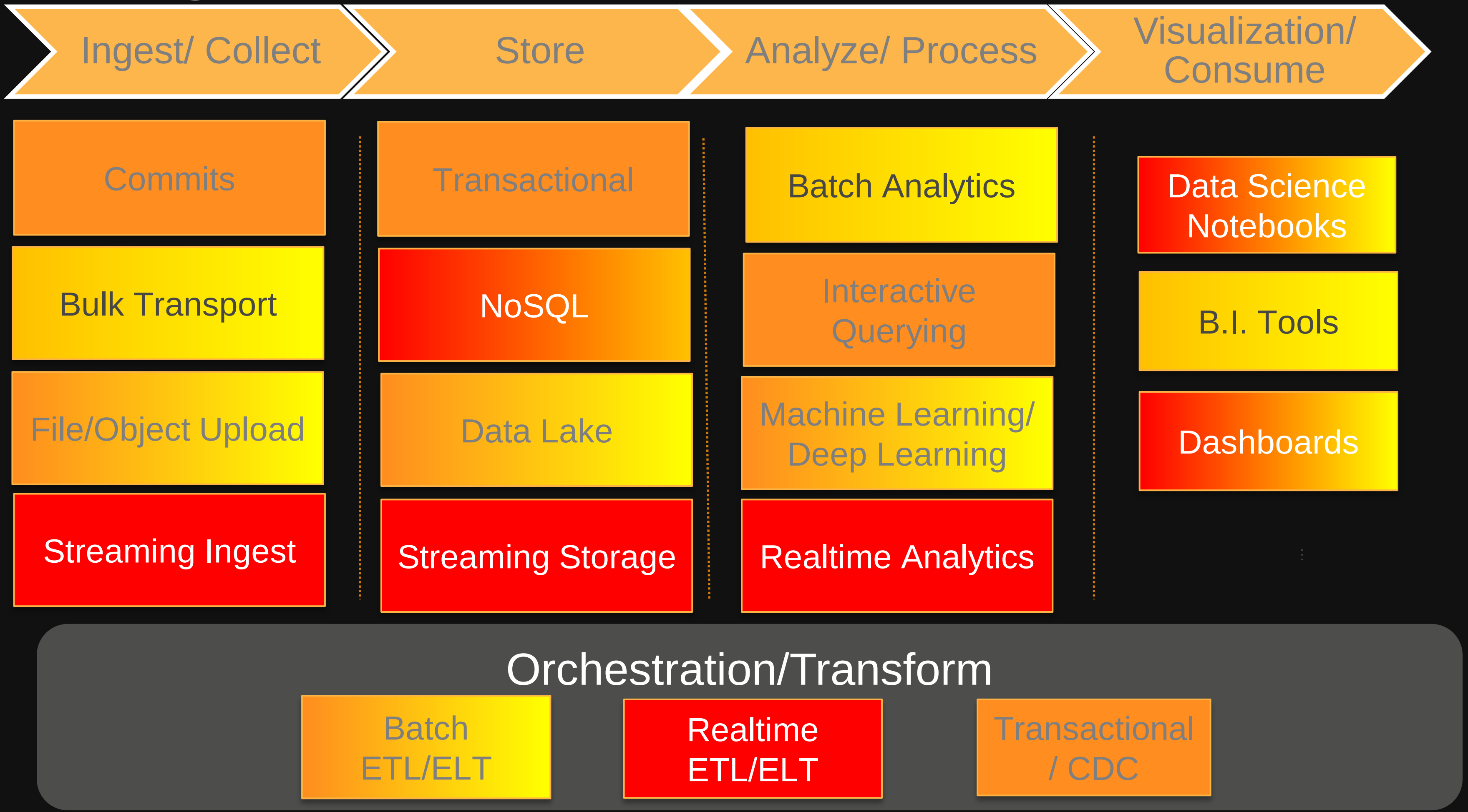
answers

Orchestrate/Transform

What Is the temperature of your data / access ?



AWS Big Data services



Ingest/ Collect

- Commits
- Bulk Transport
- File/Object Upload
- Streaming Ingest

Store

- Transactional
- NoSQL
- Data Lake
- Streaming Storage

Analyze/ Process

- Batch Analytics
- Interactive Querying
- Machine Learning/ Deep Learning
- Realtime Analytics

Visualization/ Consume

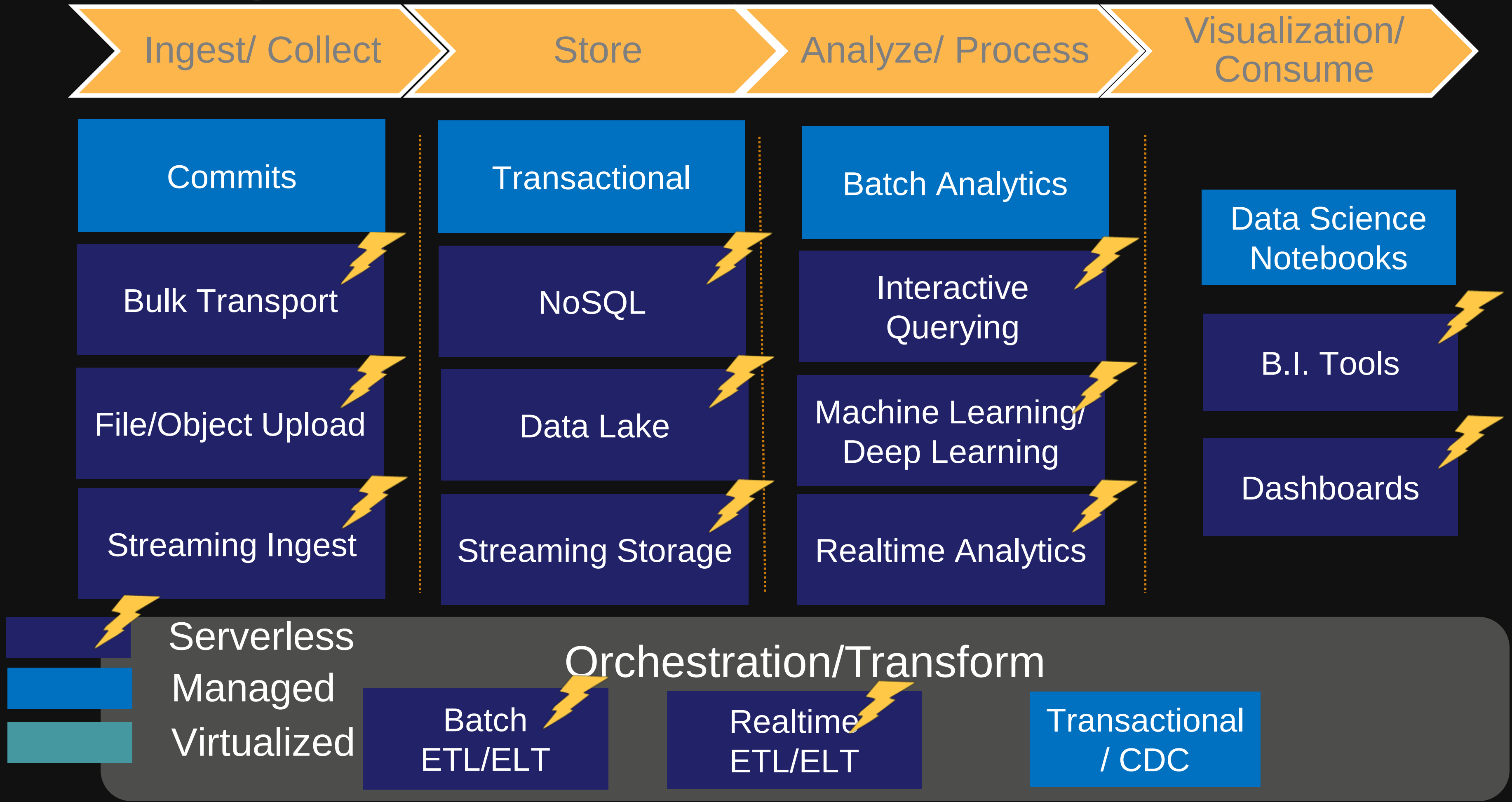
- Data Science Notebooks
- B.I. Tools
- Dashboards

Orchestration/Transform

- Batch ETL/ELT
- Realtime ETL/ELT
- Transactional / CDC

AWS Big Data services

⚡ = Serverless



Ingest/ Collect

- Commits
- Bulk Transport ⚡
- File/Object Upload ⚡
- Streaming Ingest ⚡

Store

- Transactional
- NoSQL ⚡
- Data Lake ⚡
- Streaming Storage ⚡

Analyze/ Process

- Batch Analytics
- Interactive Querying ⚡
- Machine Learning/ Deep Learning ⚡
- Realtime Analytics ⚡

Visualization/ Consume

- Data Science Notebooks
- B.I. Tools ⚡
- Dashboards ⚡

Orchestration/Transform

- Batch ETL/ELT ⚡
- Realtime ETL/ELT ⚡
- Transactional / CDC

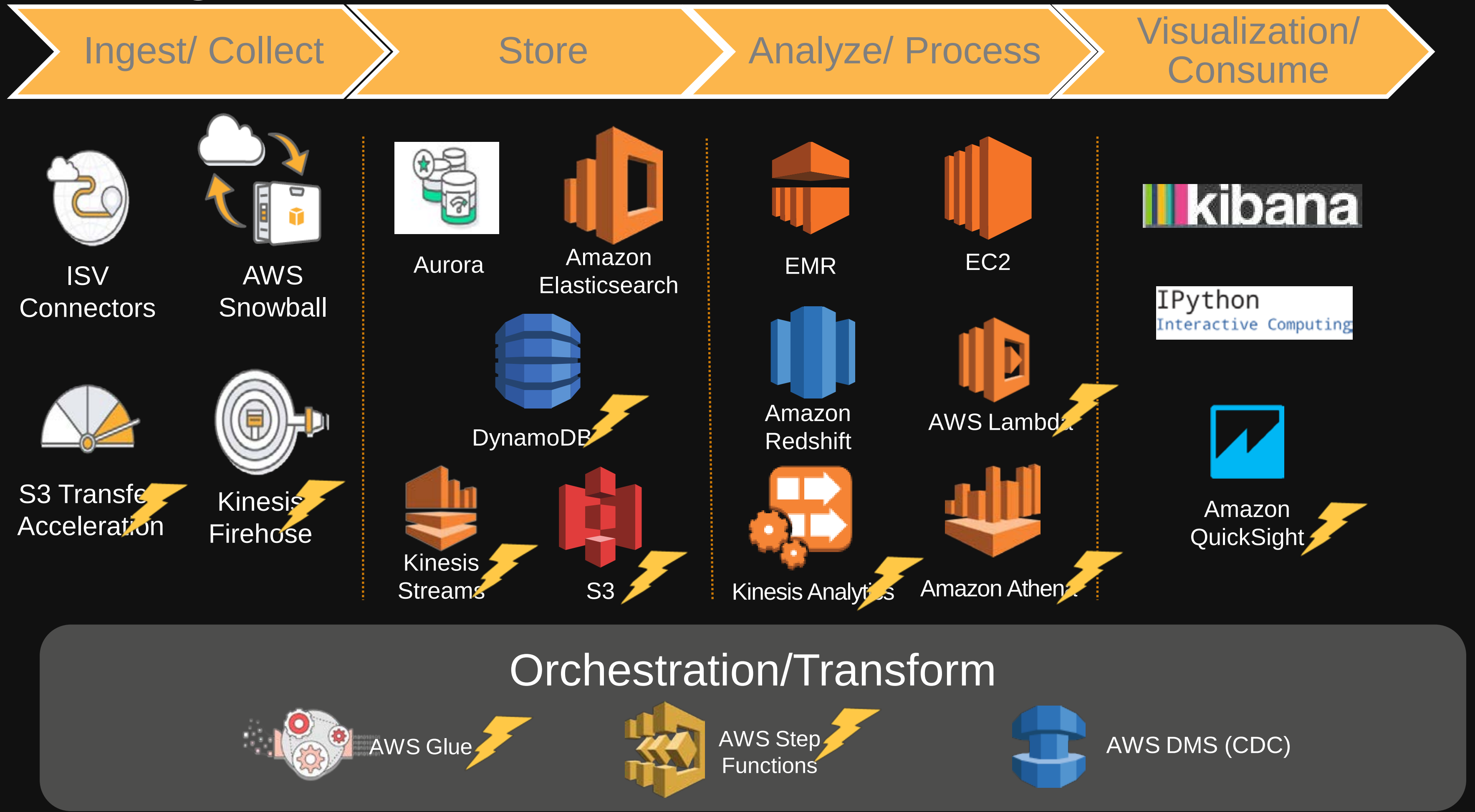
⚡ Serverless

Managed

Virtualized

AWS Big Data services

⚡ = Serverless



Ingest/ Collect

- ISV Connectors
- AWS Snowball
- S3 Transfer Acceleration ⚡
- Kinesis Firehose ⚡

Store

- Aurora
- Amazon Elasticsearch
- DynamoDB ⚡
- Kinesis Streams ⚡
- S3 ⚡

Analyze/ Process

- EMR
- EC2
- Amazon Redshift
- AWS Lambda ⚡
- Kinesis Analytics ⚡
- Amazon Athena ⚡

Visualization/ Consume

- kibana
- IPython Interactive Computing
- Amazon QuickSight ⚡

Orchestration/Transform

- AWS Glue ⚡
- AWS Step Functions ⚡
- AWS DMS (CDC)

Key Services Overview

Big Data storage for virtually all AWS services

Amazon S3

- Store anything
- Object storage
- Scalable
- 99.999999999% durability
- Extremely low cost

Fast & flexible NoSQL database service

Amazon
DynamoDB

- NoSQL database
- Seamless scalability
- Zero admin
- Single digit millisecond latency

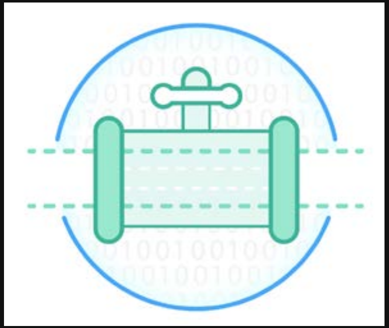
Real-time streaming platform

Amazon Kinesis

- Streams, Firehose, Analytics
- Real-time processing
- High throughput, elastic
- Easy to use
- Integration with S3, EMR, Amazon Redshift, Amazon DynamoDB

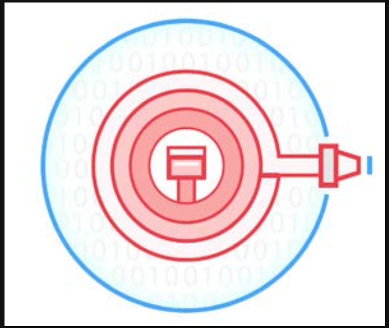
Amazon Kinesis: Streaming data made easy

Services make it easy to capture, deliver, and process streams on AWS



Amazon Kinesis Streams

- For technical developers
- Build your own custom applications that process or analyze streaming data



Amazon Kinesis Firehose

- For all developers, data scientists
- Easily load massive volumes of streaming data into S3, Amazon Redshift and Amazon Elasticsearch Service

Amazon Kinesis Analytics

- For all developers, data scientists
- Easily analyze data streams using standard SQL queries

Serverless compute

AWS Lambda

- Run your code in the cloud - fully managed and highly available
- Triggered through API or state changes in your setup
- Scales automatically to match the incoming event rate
- Node.js (JavaScript), Python, Java, and C#
- Charged per 100ms execution time

Interactive query service

Amazon
Athena

- Query directly from Amazon S3
- Use ANSI SQL
- Serverless
- Multiple data formats
- Pay per query

Fully managed ETL service

In Preview

AWS Glue

- Catalog data sources
- Identify data formats & data types
- Error handling
- Manage and scale resources
- Generate ETL code
- Schedules & executes ETL jobs

AWS Glue: Services

In Preview

Data catalog

- Hive metastore-compatible metadata repository of data sources.
- Crawls data source to infer table, data type, partition format.

Job authoring

- Generates Python code to move data from source to destination.
- Edit with your favorite IDE; share code snippets using Git.

Job execution

- Runs jobs in Spark containers – automatic scaling based on SLA.
- AWS Glue is serverless – only pay for the resources you consume.

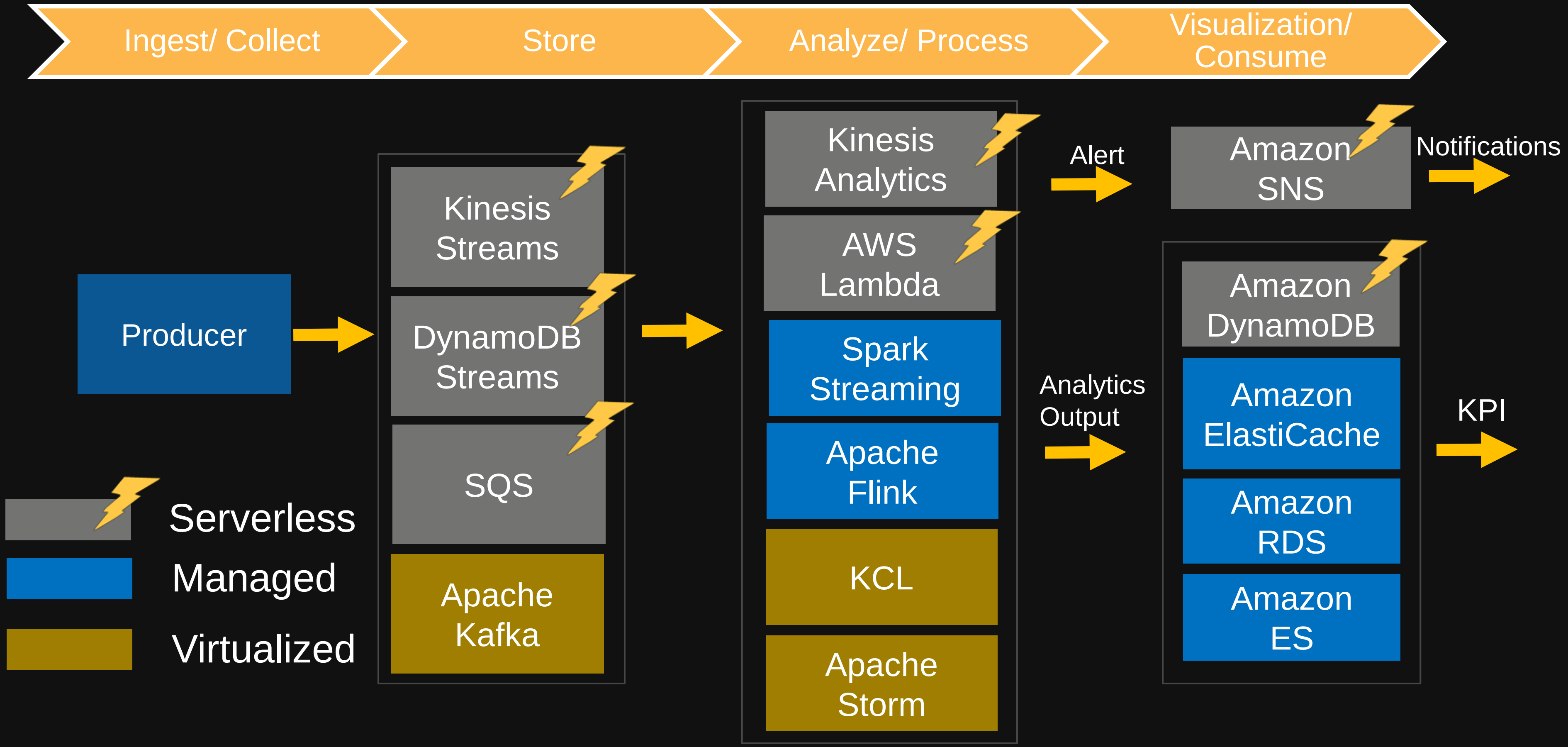
Business Intelligence

Amazon QuickSight

- Fast and cloud-powered
- Easy to use, no infrastructure to manage
- Scales to hundreds of thousands of users
- Quick calculations with SPICE
- 1/10th the cost of legacy BI software

Serverless design patterns

Real-time analytics



Ingest/ Collect

Store

Analyze/ Process

Visualization/ Consume

Producer

Kinesis Streams

DynamoDB Streams

SQS

Apache Kafka

Kinesis Analytics

AWS Lambda

Spark Streaming

Apache Flink

KCL

Apache Storm

Alert

Analytics Output

Amazon SNS

Notifications

Amazon DynamoDB

Amazon ElastiCache

Amazon RDS

Amazon ES

KPI

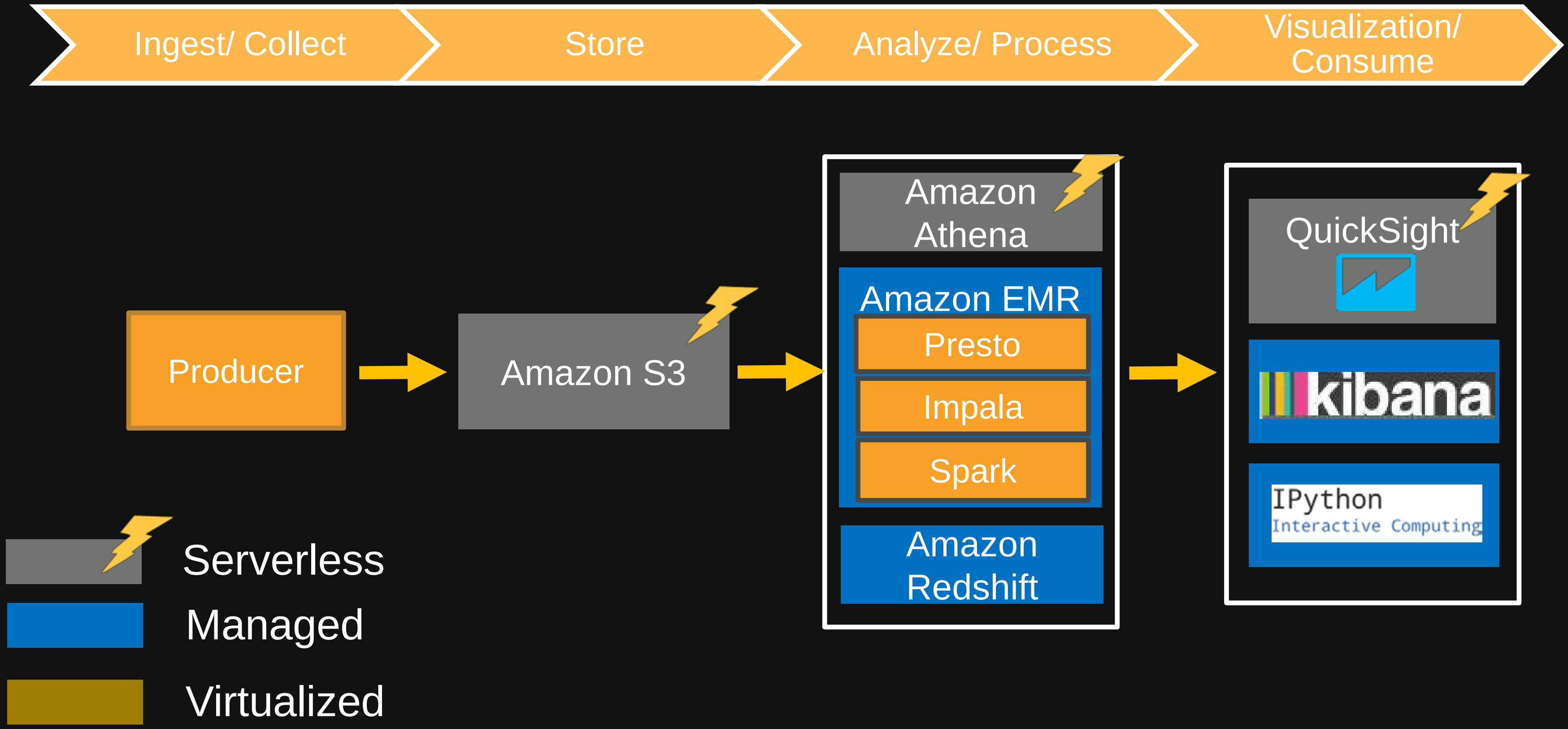
Serverless

Managed

Virtualized

Interactive Queries

Interactive



Ingest/ Collect

Store

Analyze/ Process

Visualization/ Consume

Producer

Amazon S3

Amazon Athena

Amazon EMR

Presto

Impala

Spark

Amazon Redshift

QuickSight

kibana

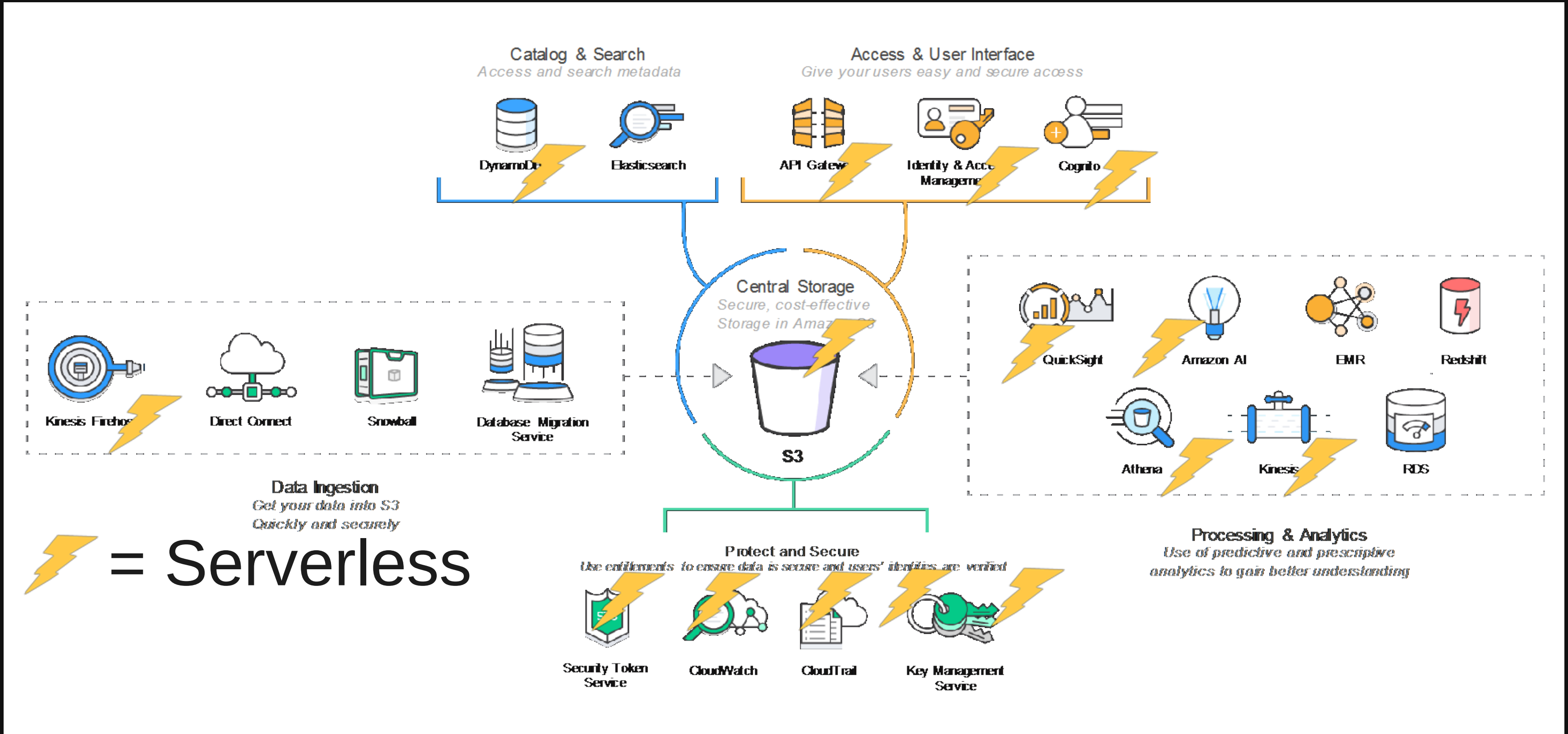
IPython Interactive Computing

Serverless

Managed

Virtualized

Data lake reference architecture



Catalog & Search

Access and search metadata

DynamoDı

Elasticsearch

Access & User Interface

Give your users easy and secure access

API Gatew

Identity & Acce Manageme

Cognito

Central Storage

Secure, cost-effective Storage in Amaz S3

S3

Kinesis Firehos

Direct Connect

Snowball

Database Migration Service

Data Ingestion

Get your data into S3 Quickly and securely

QuickSight

Amazon AI

EMR

Redshift

Athena

Kinesis

RDS

Processing & Analytics

Use of predictive and prescriptive analytics to gain better understanding

Protect and Secure

Use entitlements to ensure data is secure and users' identities are verified

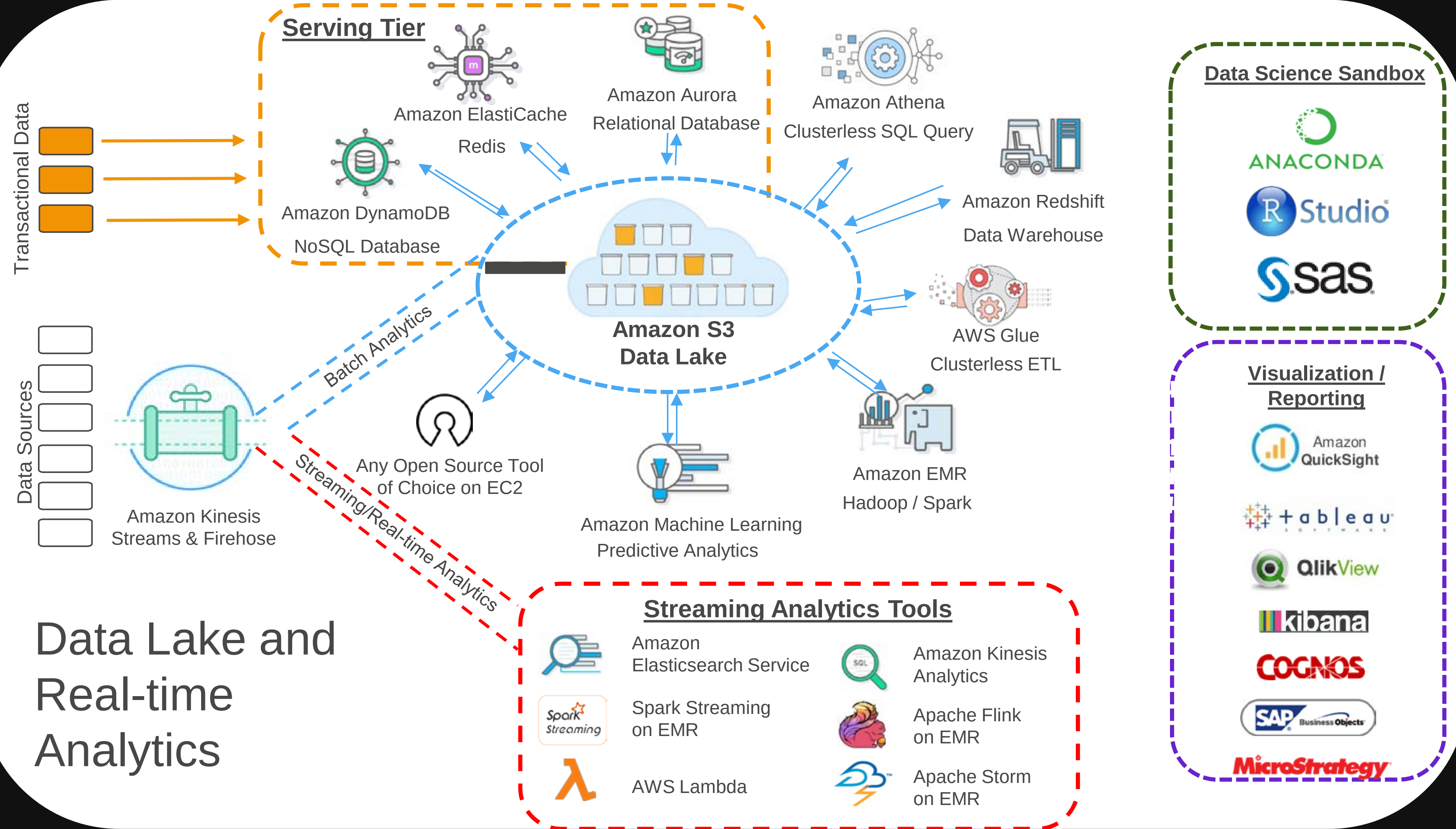
Security Token Service

CloudWatch

CloudTrail

Key Management Service

= Serverless



Transactional Data

Serving Tier

Amazon ElastiCache
Redis

Amazon Aurora
Relational Database

Amazon DynamoDB
NoSQL Database

[redacted]

Amazon Athena
Clusterless SQL Query

Amazon Redshift
Data Warehouse

Amazon S3
Data Lake

AWS Glue
Clusterless ETL

Batch Analytics

Any Open Source Tool
of Choice on EC2

Amazon Machine Learning
Predictive Analytics

Amazon EMR
Hadoop / Spark

Data Sources

Amazon Kinesis
Streams & Firehose

Streaming/Real-time Analytics

Streaming Analytics Tools

Amazon
Elasticsearch Service

Amazon Kinesis
Analytics

Spark Streaming
on EMR

Apache Flink
on EMR

AWS Lambda

Apache Storm
on EMR

Data Science Sandbox

ANACONDA

RStudio

sas

Visualization / Reporting

Amazon
QuickSight

tableau

QlikView

kibana

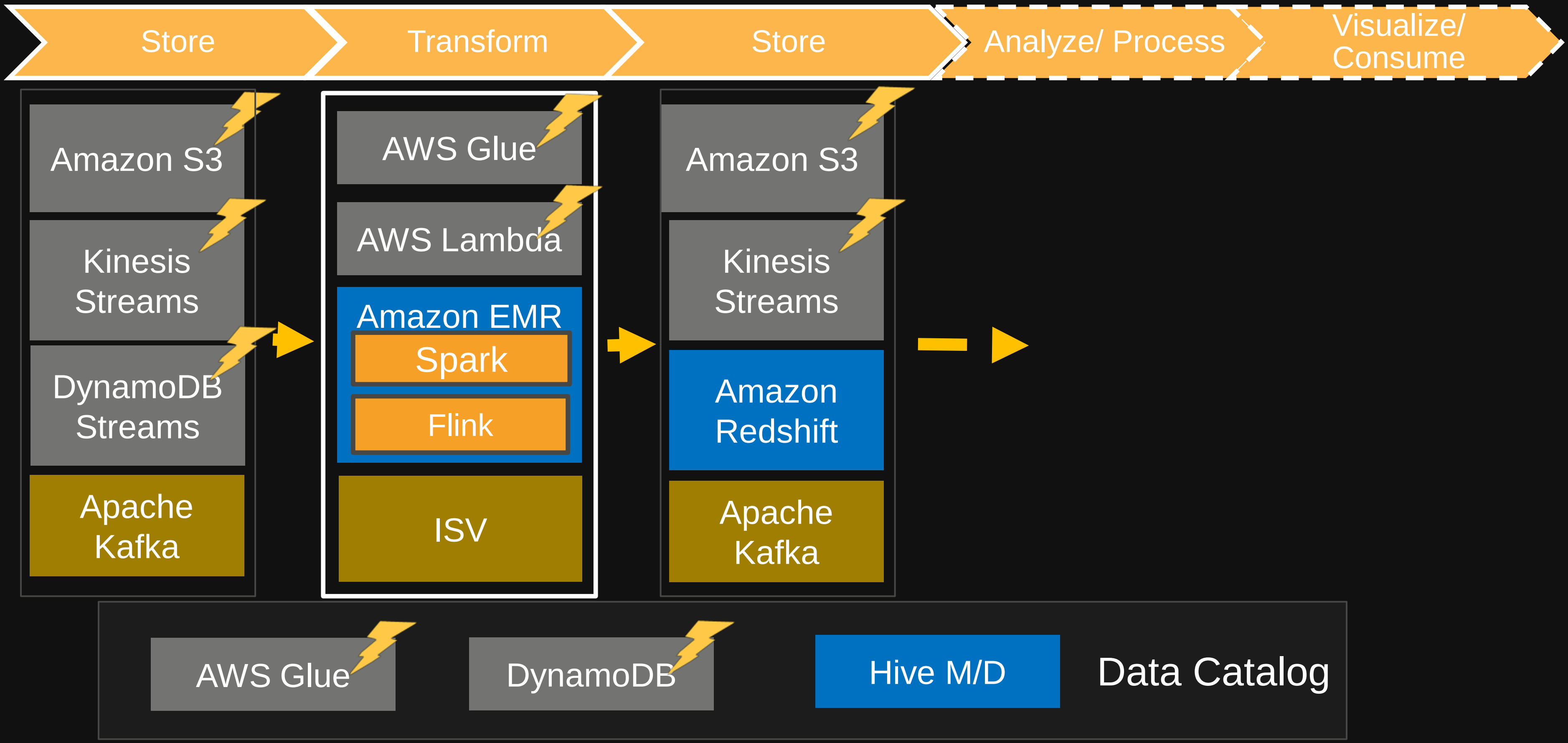
COGNOS

SAP Business Objects

MicroStrategy

Data Lake and Real-time Analytics

Serverless ETL



Serverless nicely fits into big data platforms

- AWS serverless Big Data services
  - Complements existing big data flows
  - Focus on the analytics and not on infrastructure or servers
  - Don’t focus on the scaling, availability, and undifferentiated heavy lifting

- Pay only for what you use
- Easily try out different tools, analytics, and solutions

AWS

SUMMIT

Thank you!

amazon web services