AWS

SUMMIT

Building a Data Processing Pipeline on AWS

Unni Pillai

Solutions Architect

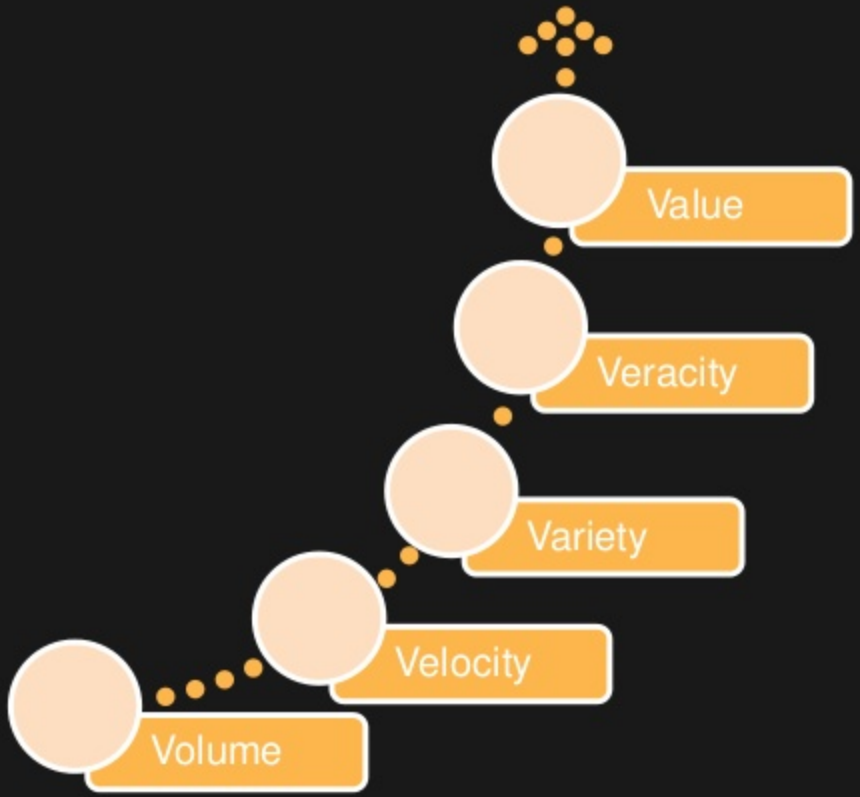
unni_k_pillai

amazon web services

Agenda

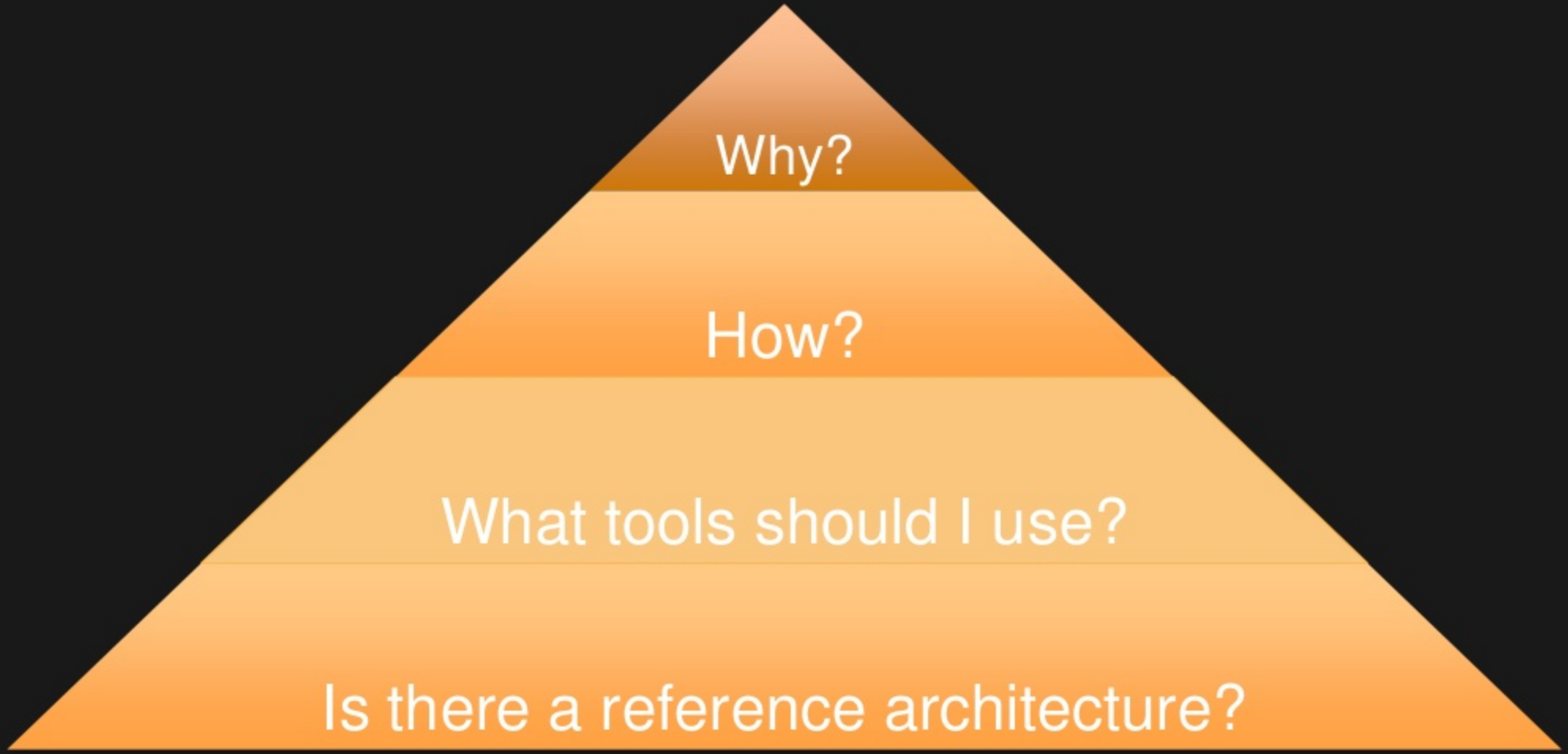
- Big Data Challenges
- Architectural Principles
- Stages in a Data Processing Pipeline
- Demo : Build a data processing pipeline
- Design Patterns

Ever Increasing Big Data



Plethora of Tools

Big Data Challenges



Architectural Principles

Build **decoupled** systems

- Data → Store → Process → Store → Analyze → Answers

Use the **right tool** for the job

- Data structure, latency, throughput, access patterns

Leverage AWS **managed services**

- Scalable/elastic, available, reliable, secure, no/low admin

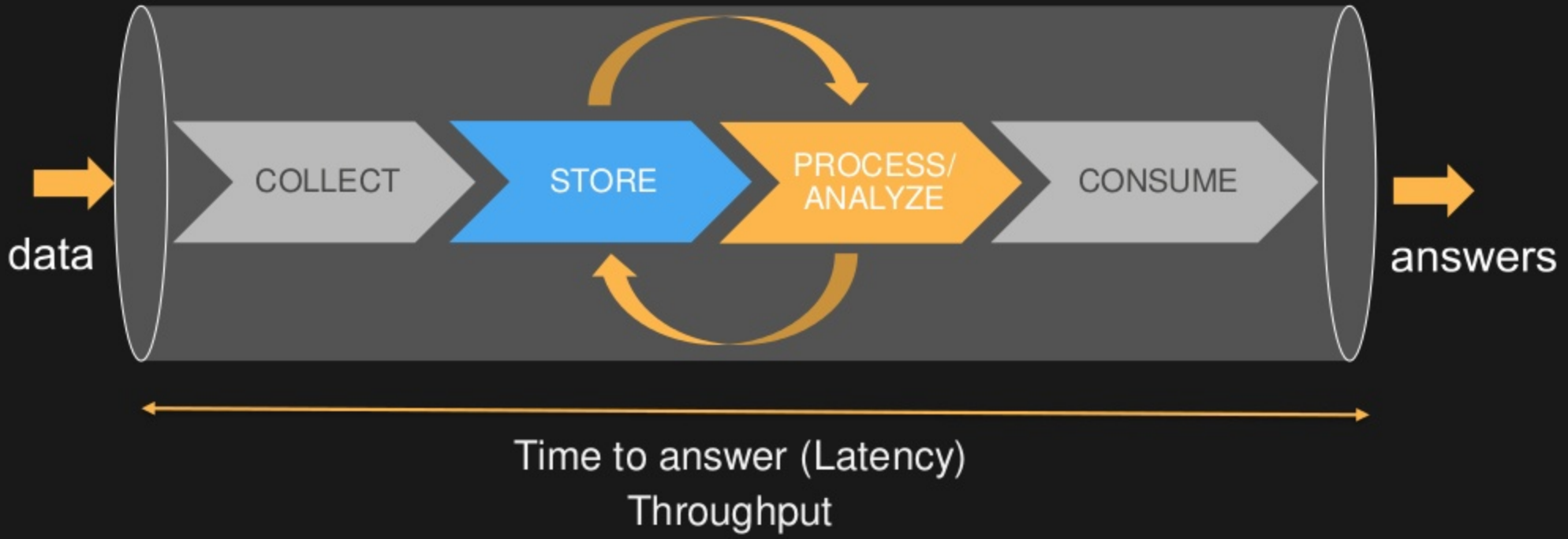
Use **log-centric** design patterns

- Immutable logs, materialized views

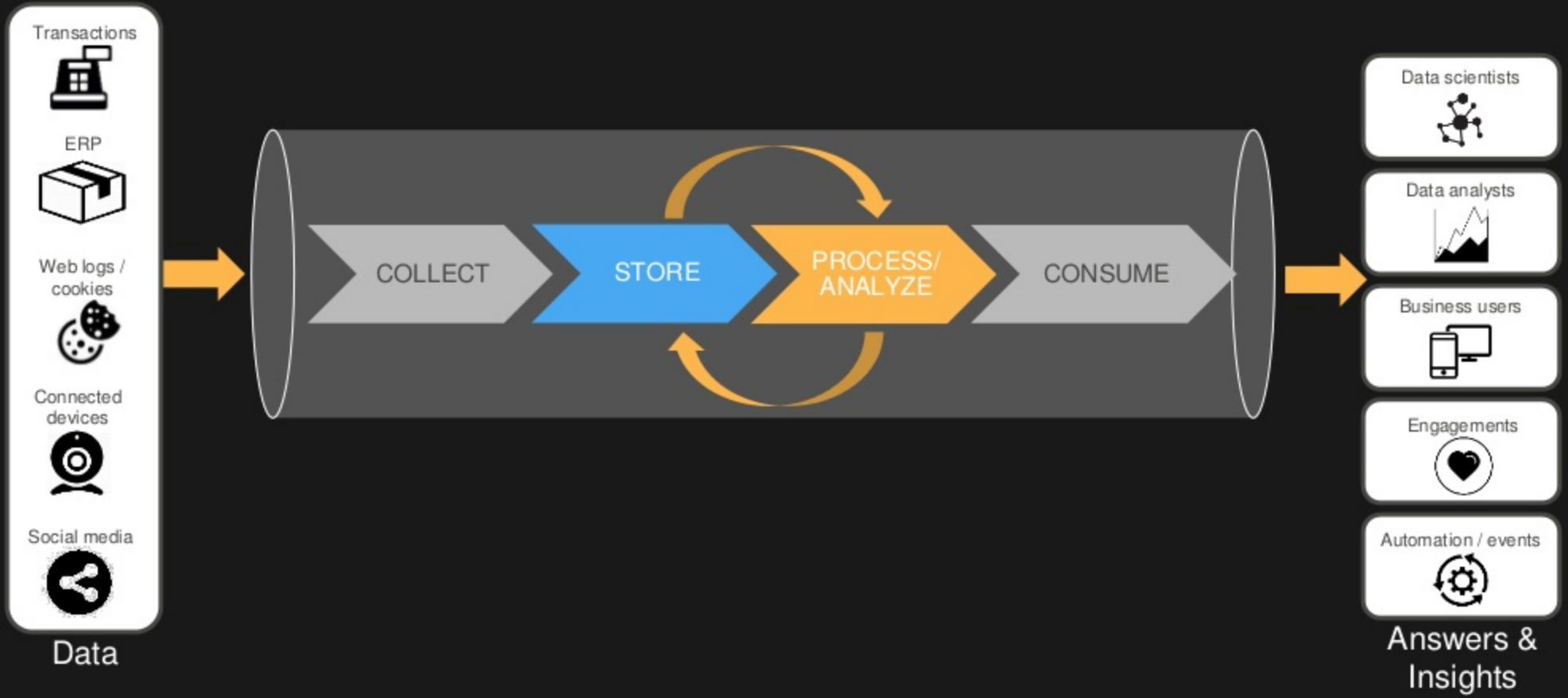
Be **cost-conscious**

- Big data ≠ big cost

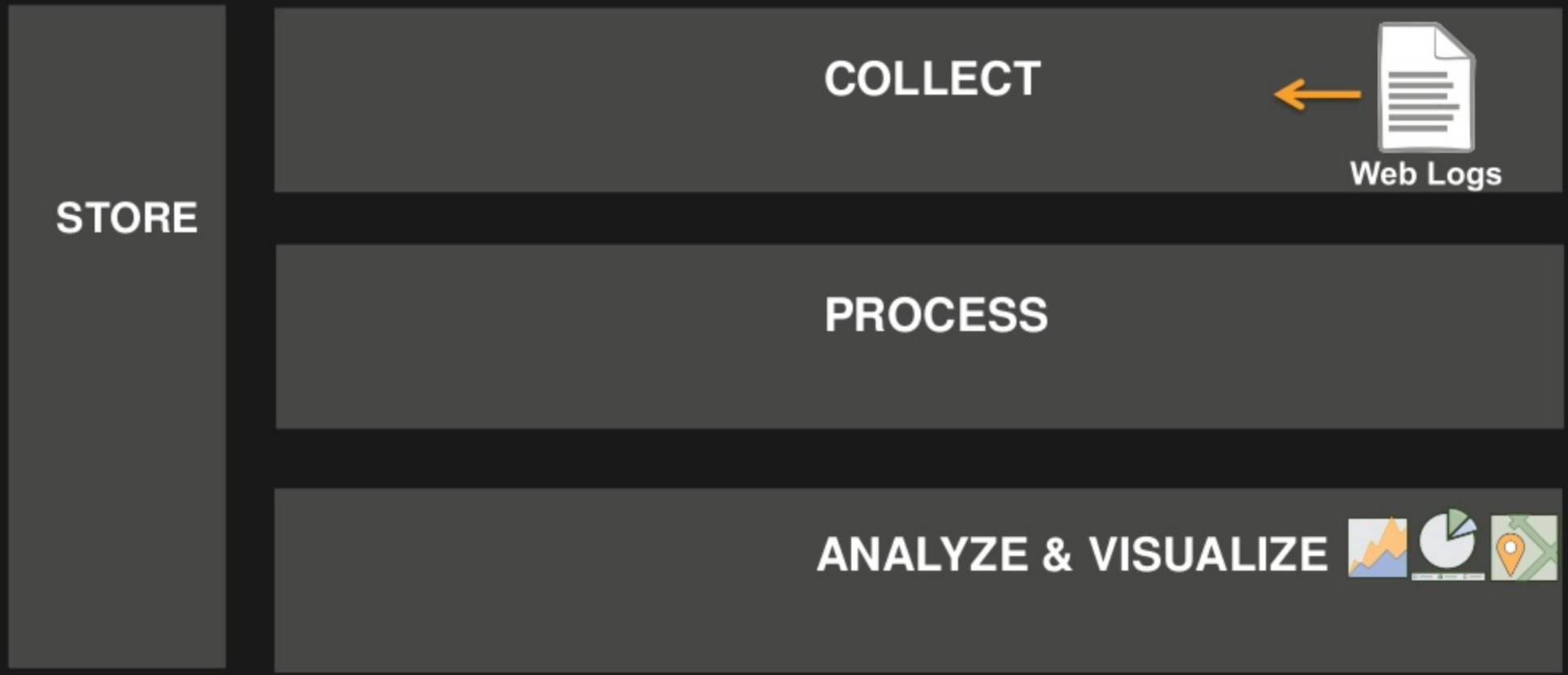
Simplify Big Data Processing



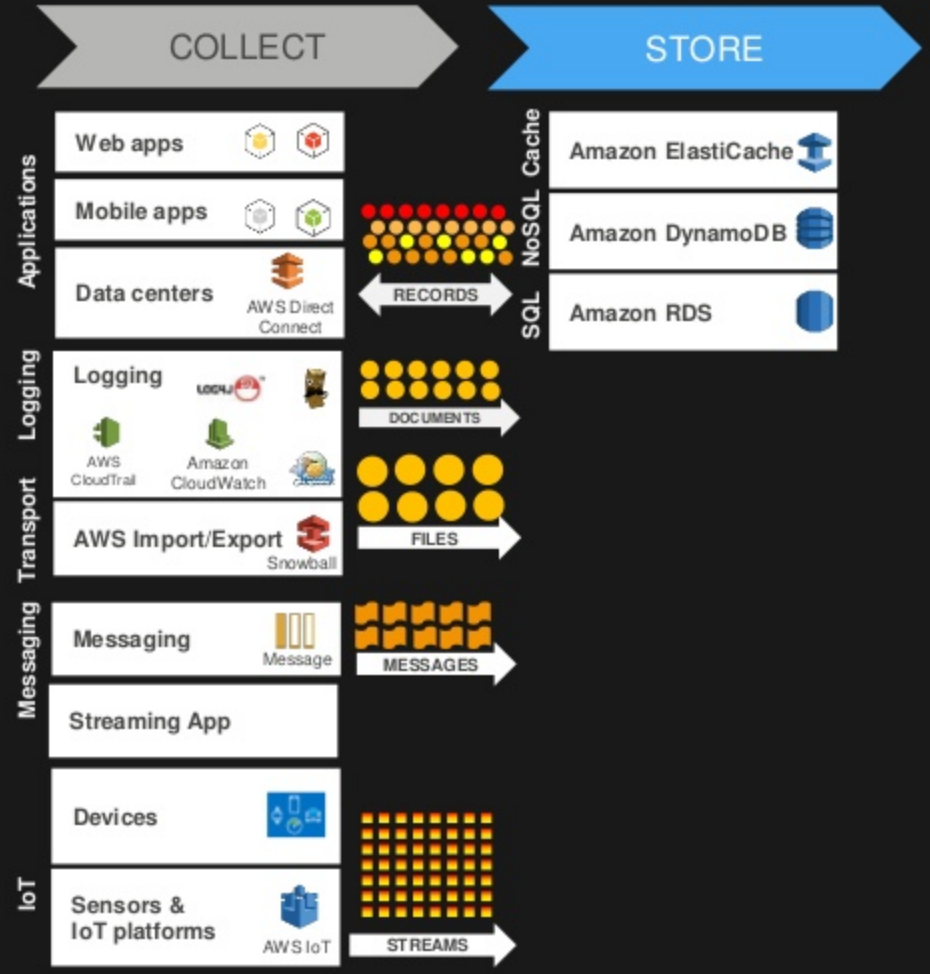
Simplify Big Data Processing

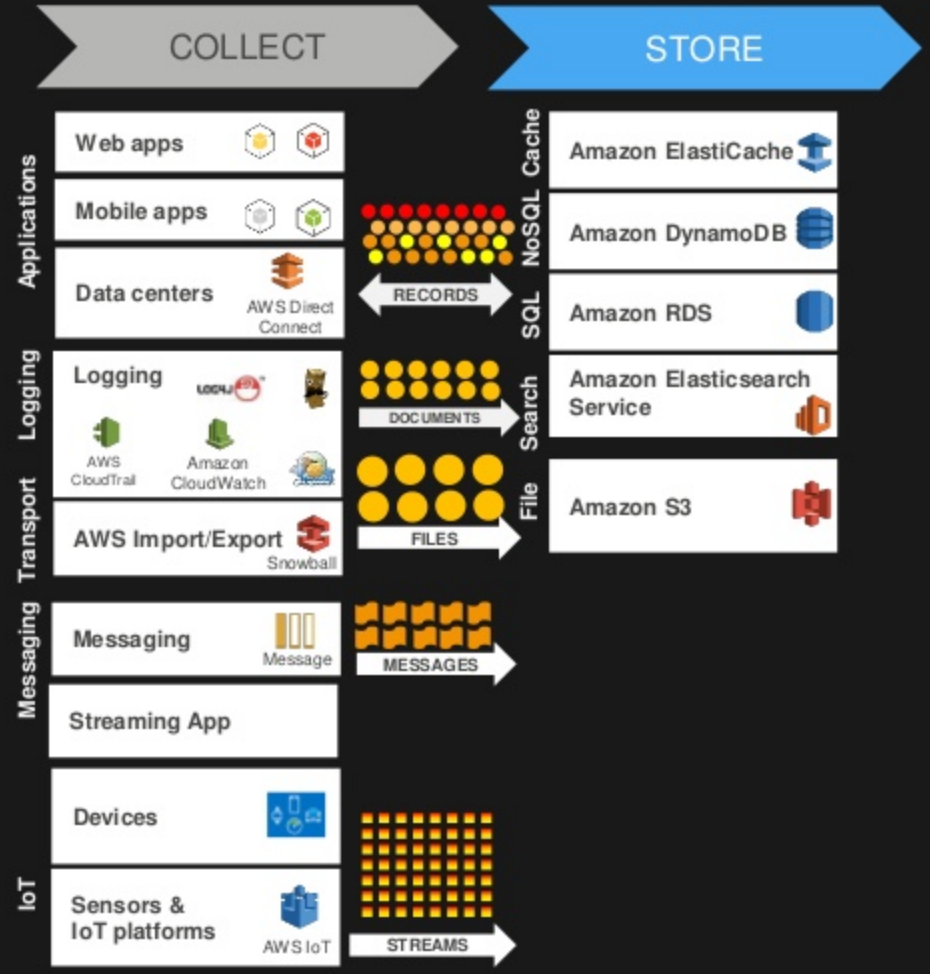


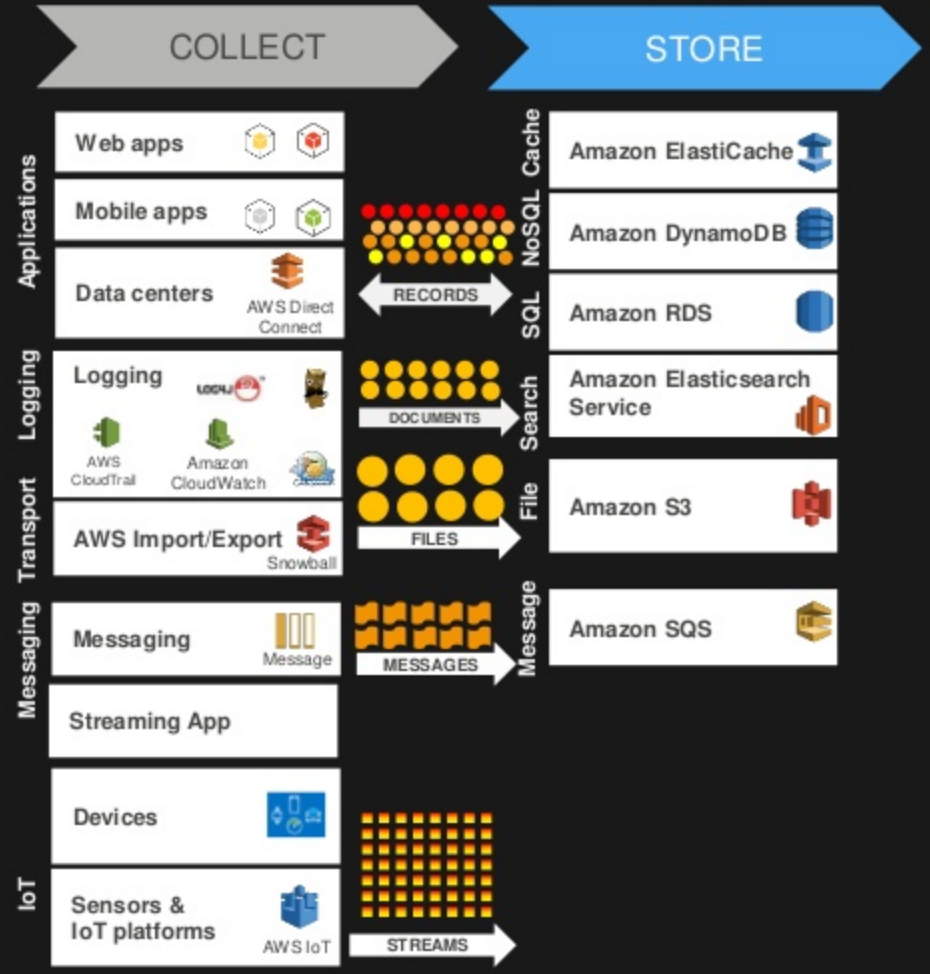
Building a pipeline - DEMO

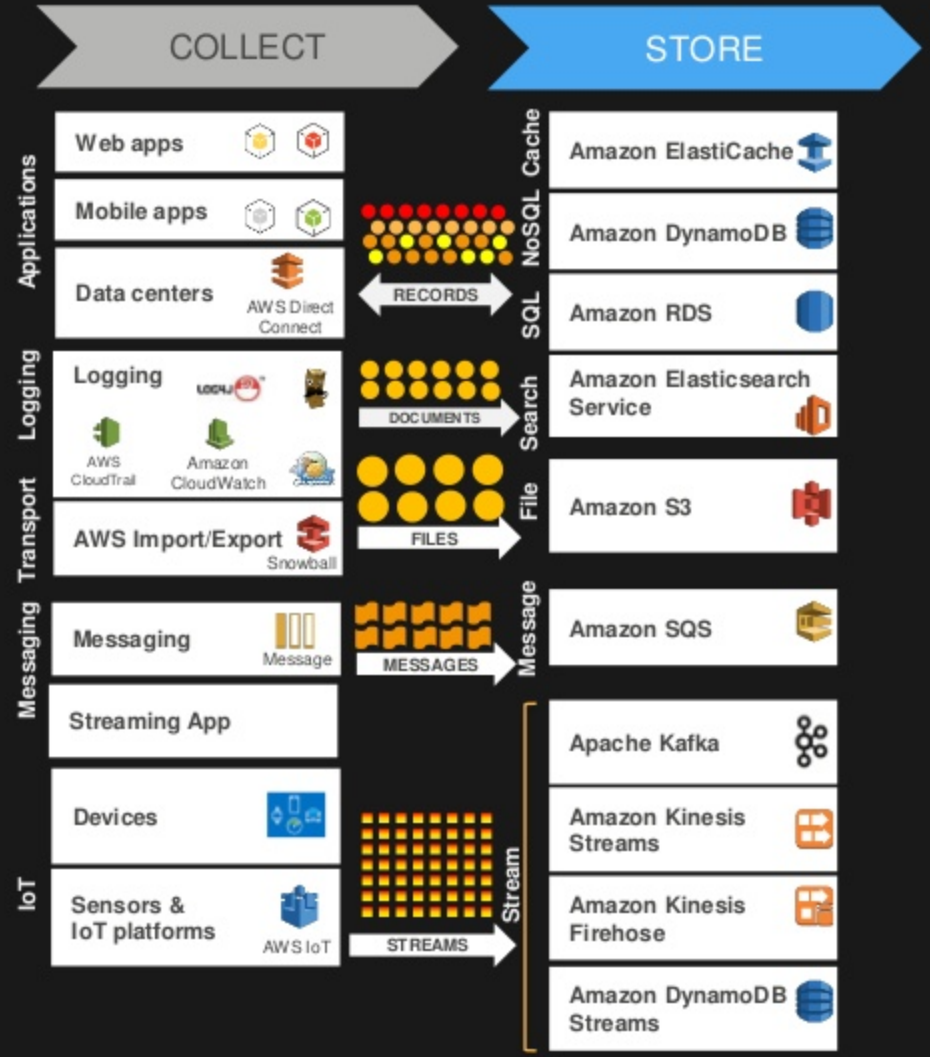


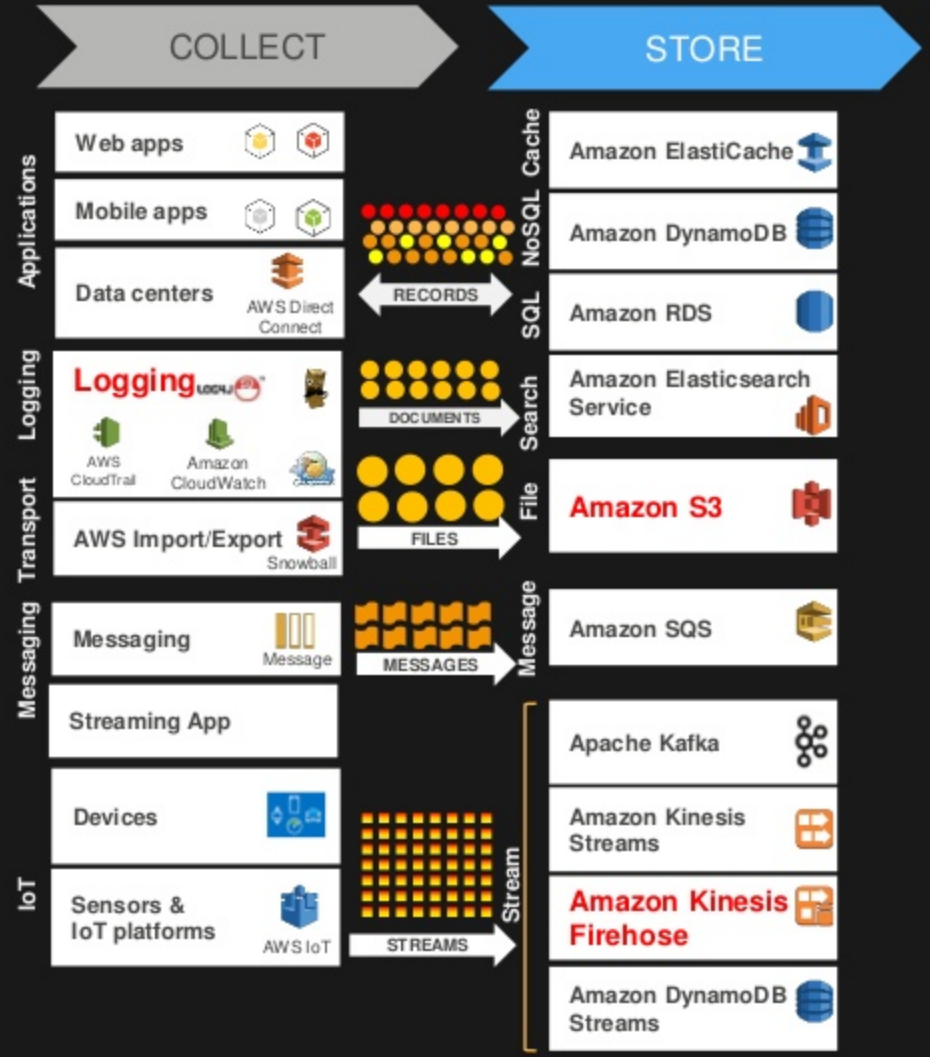

COLLECT
STORE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
RECORDS
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
DOCUMENTS
Transport
AWS Import/Export
Snowball
FILES
Messaging
Messaging
Message
MESSAGES
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
STREAMS


COLLECT
STORE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS


COLLECT
STORE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3

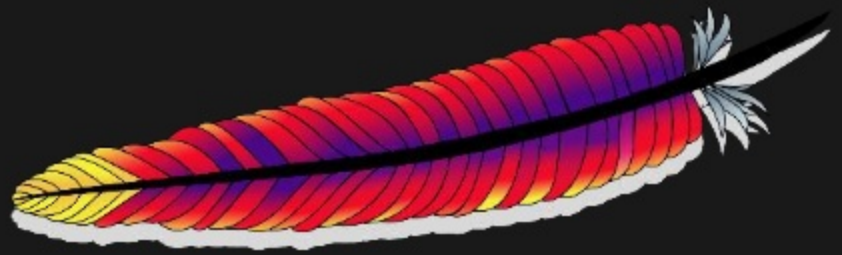

COLLECT
STORE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS


COLLECT
STORE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Stream
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams


COLLECT
STORE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Stream
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams

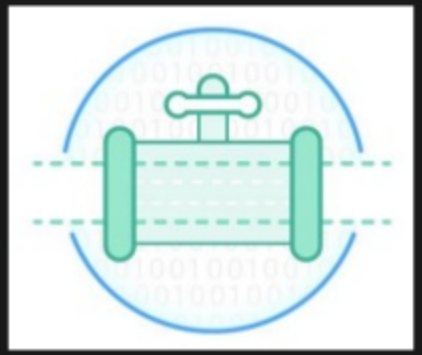
Weblogs – Common Log Format (CLF)

```
75.35.230.210 - - [20/Jul/2016:22:22:42 -0700]
"GET /images/pigtrihawk.jpg HTTP/1.1" 200 29236
"http://www.swivel.com/graphs/show/1163466"
"Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.11)
Gecko/2009060215 Firefox/3.0.11 (.NET CLR 3.5.30729)"
```



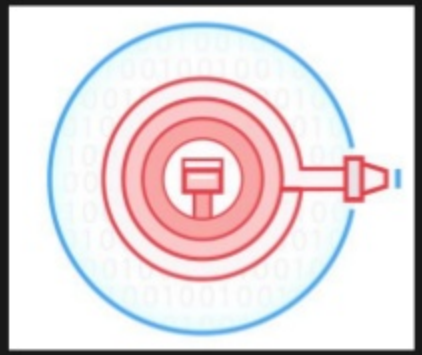
Amazon Kinesis: Streaming Data Made Easy

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

Amazon Kinesis Analytics

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

Why Is Amazon S3 Good for Big Data?

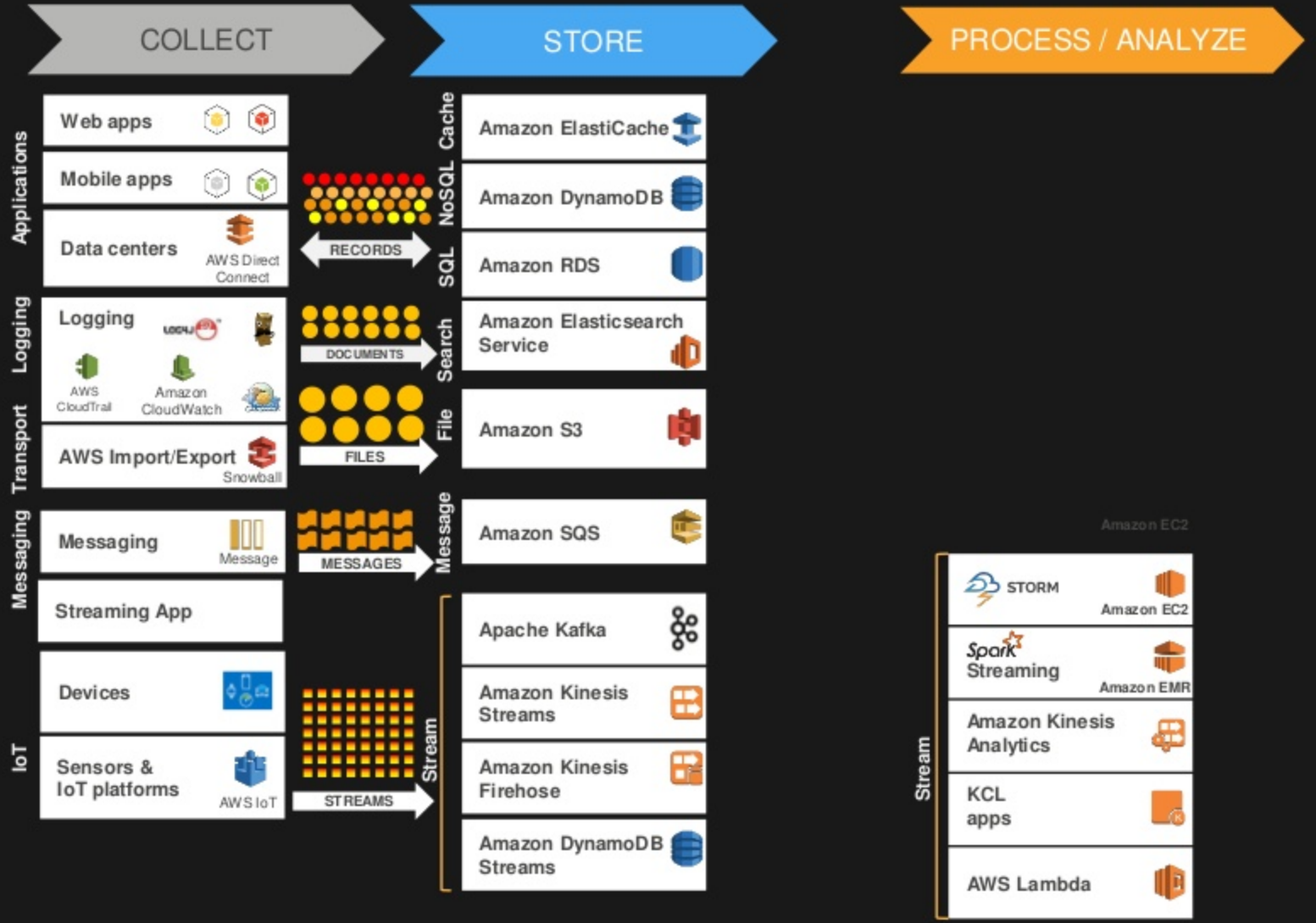
- **Unlimited** number of objects and **volume** of data
- Very high bandwidth – no aggregate throughput limit
- Natively supported by big data frameworks (**Spark**, Hive, Presto, etc.)
- No need to run compute clusters for storage (unlike HDFS)
- Multiple & heterogeneous analysis clusters can use the same data
- Designed for **99.99% availability** – can tolerate zone failure
- Designed for **99.999999999% durability**
- No need to pay for data replication
- Native support for versioning
- Tiered-storage (Standard, IA, Amazon Glacier) via life-cycle policies
- **Secure** – SSL, client/server-side encryption at rest
- **Low cost**

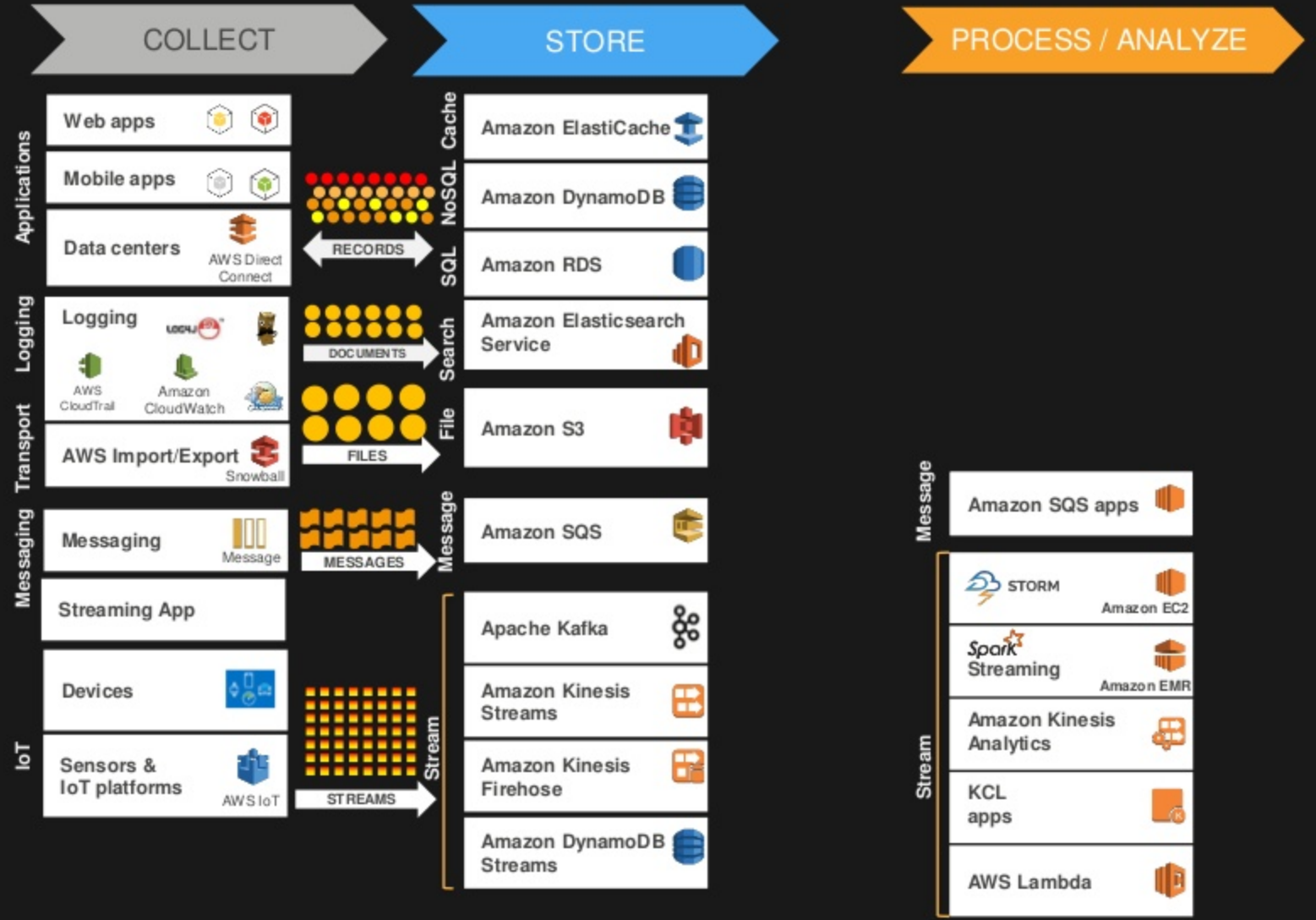
Building a pipeline - DEMO

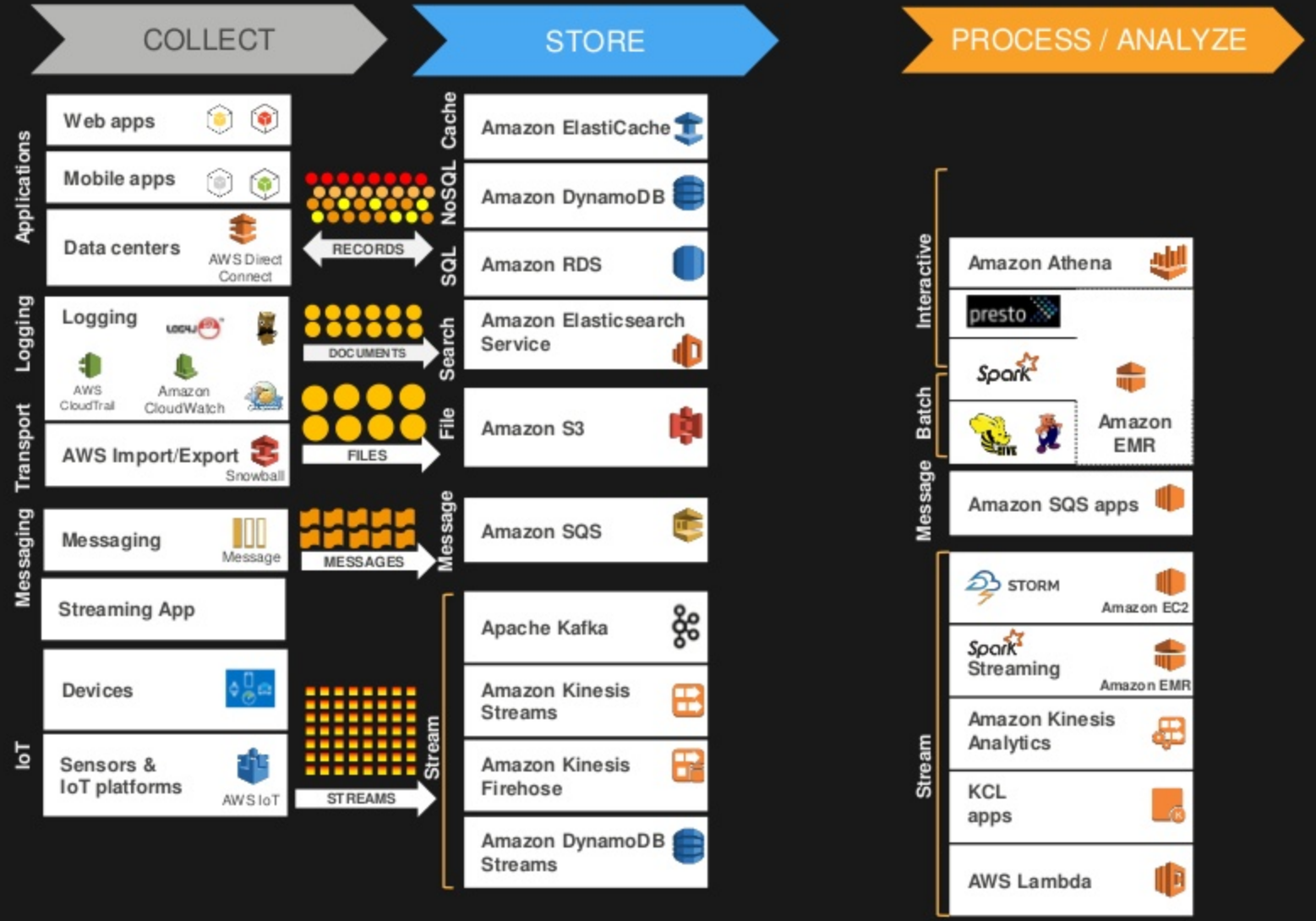
Demo

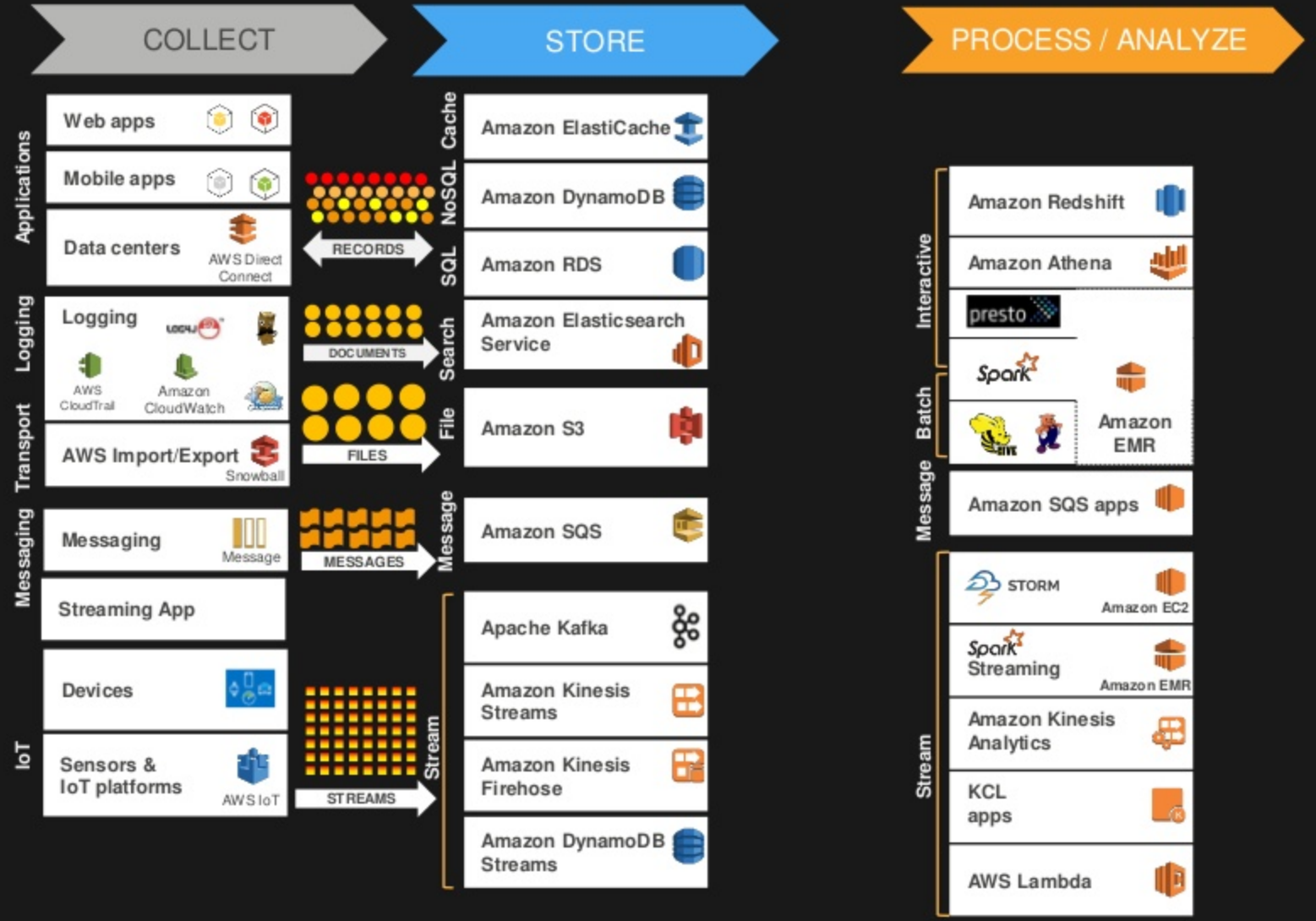
- Create Amazon S3 Bucket
- Create Amazon Kinesis Firehose delivery stream
- Publish logs to Amazon Kinesis Firehose

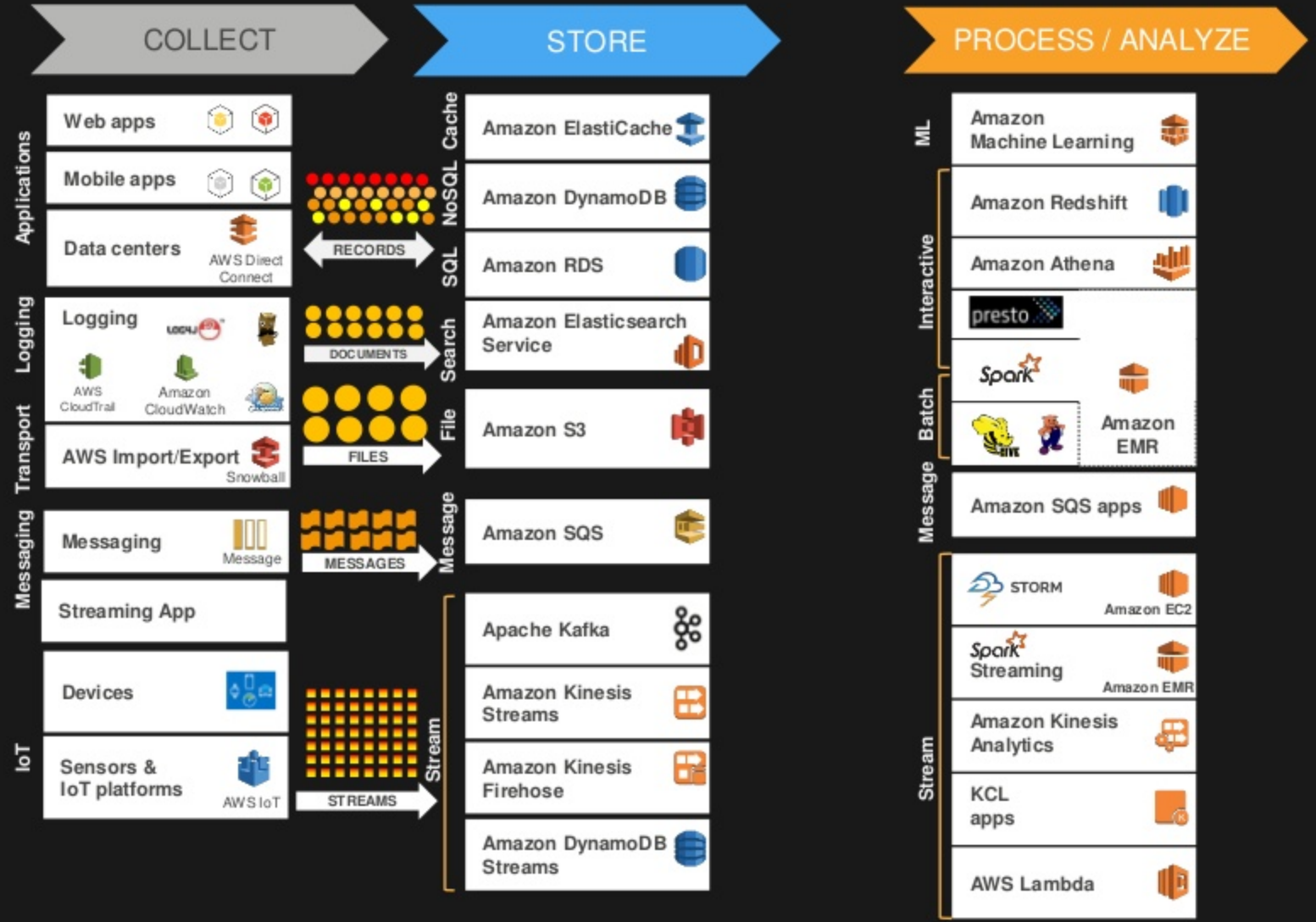

COLLECT
STORE
PROCESS / ANALYZE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Stream
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams

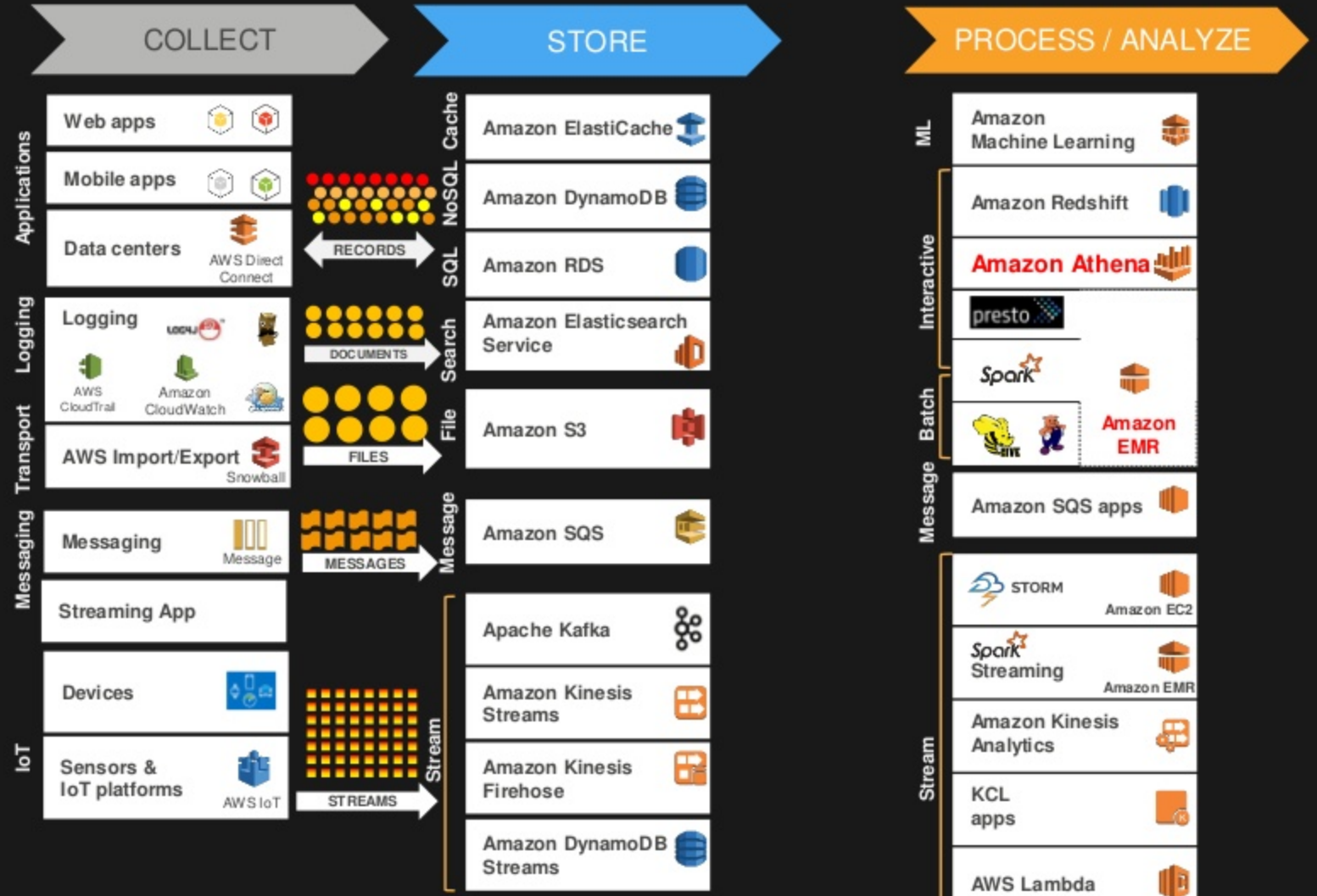

COLLECT
STORE
PROCESS / ANALYZE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Stream
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams
Amazon EC2
Stream
STORM
Amazon EC2
Spark Streaming
Amazon EMR
Amazon Kinesis Analytics
KCL apps
AWS Lambda


COLLECT
STORE
PROCESS / ANALYZE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Stream
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams
Message
Amazon SQS apps
Stream
STORM
Amazon EC2
Spark Streaming
Amazon EMR
Amazon Kinesis Analytics
KCL apps
AWS Lambda


COLLECT
STORE
PROCESS / ANALYZE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Stream
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams
Interactive
Batch
Message
Stream
Amazon Athena
presto
Spark
Amazon EMR
Amazon SQS apps
STORM
Amazon EC2
Spark Streaming
Amazon EMR
Amazon Kinesis Analytics
KCL apps
AWS Lambda


COLLECT
STORE
PROCESS / ANALYZE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Stream
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams
Interactive
Batch
Message
Stream
Amazon Redshift
Amazon Athena
presto
Spark
Amazon EMR
Amazon SQS apps
STORM
Amazon EC2
Spark Streaming
Amazon EMR
Amazon Kinesis Analytics
KCL apps
AWS Lambda


COLLECT
STORE
PROCESS / ANALYZE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Stream
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams
ML
Amazon Machine Learning
Interactive
Amazon Redshift
Amazon Athena
presto
Batch
Spark
Amazon EMR
Message
Amazon SQS apps
Stream
STORM
Amazon EC2
Spark Streaming
Amazon EMR
Amazon Kinesis Analytics
KCL apps
AWS Lambda


COLLECT
STORE
PROCESS / ANALYZE
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Stream
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams
ML
Amazon Machine Learning
Interactive
Amazon Redshift
Amazon Athena
presto
Batch
Spark
Amazon EMR
Message
Amazon SQS apps
Stream
STORM
Amazon EC2
Spark Streaming
Amazon EMR
Amazon Kinesis Analytics
KCL apps
AWS Lambda

Building a pipeline - DEMO


COLLECT
STORE
ETL
PROCESS / ANALYZE
CONSUME
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Stream
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams
ML
Amazon Machine Learning
Interactive
Amazon Redshift
Amazon Athena
presto
Batch
Spark
Amazon EMR
Message
Amazon SQS apps
Stream
STORM
Amazon EC2
Spark Streaming
Amazon EMR
Amazon Kinesis Analytics
KCL apps
AWS Lambda


COLLECT
STORE
ETL
PROCESS / ANALYZE
CONSUME
Applications
Web apps
Mobile apps
Data centers
AWS Direct Connect
Logging
Logging
AWS CloudTrail
Amazon CloudWatch
Transport
AWS Import/Export
Snowball
Messaging
Messaging
Message
Streaming App
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
NoSQL
SQL
Search
File
Message
Stream
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon Elasticsearch Service
Amazon S3
Amazon SQS
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams
ML
Amazon Machine Learning
Interactive
Amazon Redshift
Amazon Athena
presto
Batch
Spark
Amazon EMR
Message
Amazon SQS apps
Stream
STORM
Amazon EC2
Spark Streaming
Amazon EMR
Amazon Kinesis Analytics
KCL apps
AWS Lambda
API
Apps & Services
Analysis & visualization
Amazon QuickSight
kibana
tableau
looker
MicroStrategy
TIBCO Jaspersoft
Flot
Notebooks
Apache Zeppelin
jupyter
IDE
RStudio

Building a pipeline

Demo

- Check the files which were ingested into Amazon S3
- Clean the data using Amazon EMR (Spark)
- Create a table in Amazon Athena
    - Query data using SQL

Demo

- Amazon QuickSight Demo

Design Patterns

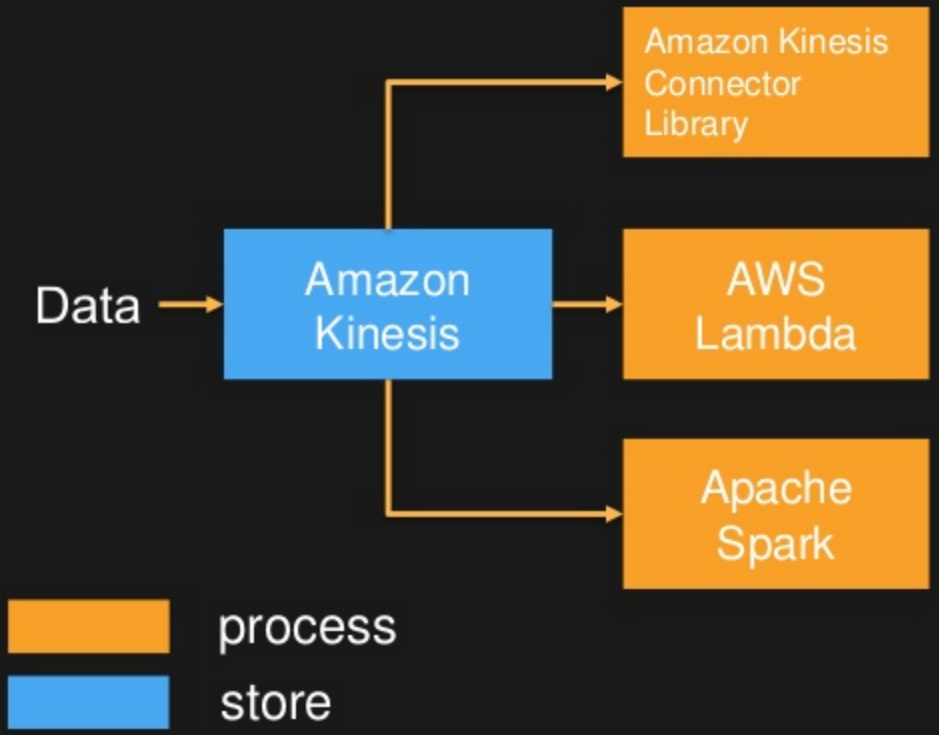
Primitive: Decoupled Data Bus

Storage decoupled from processing

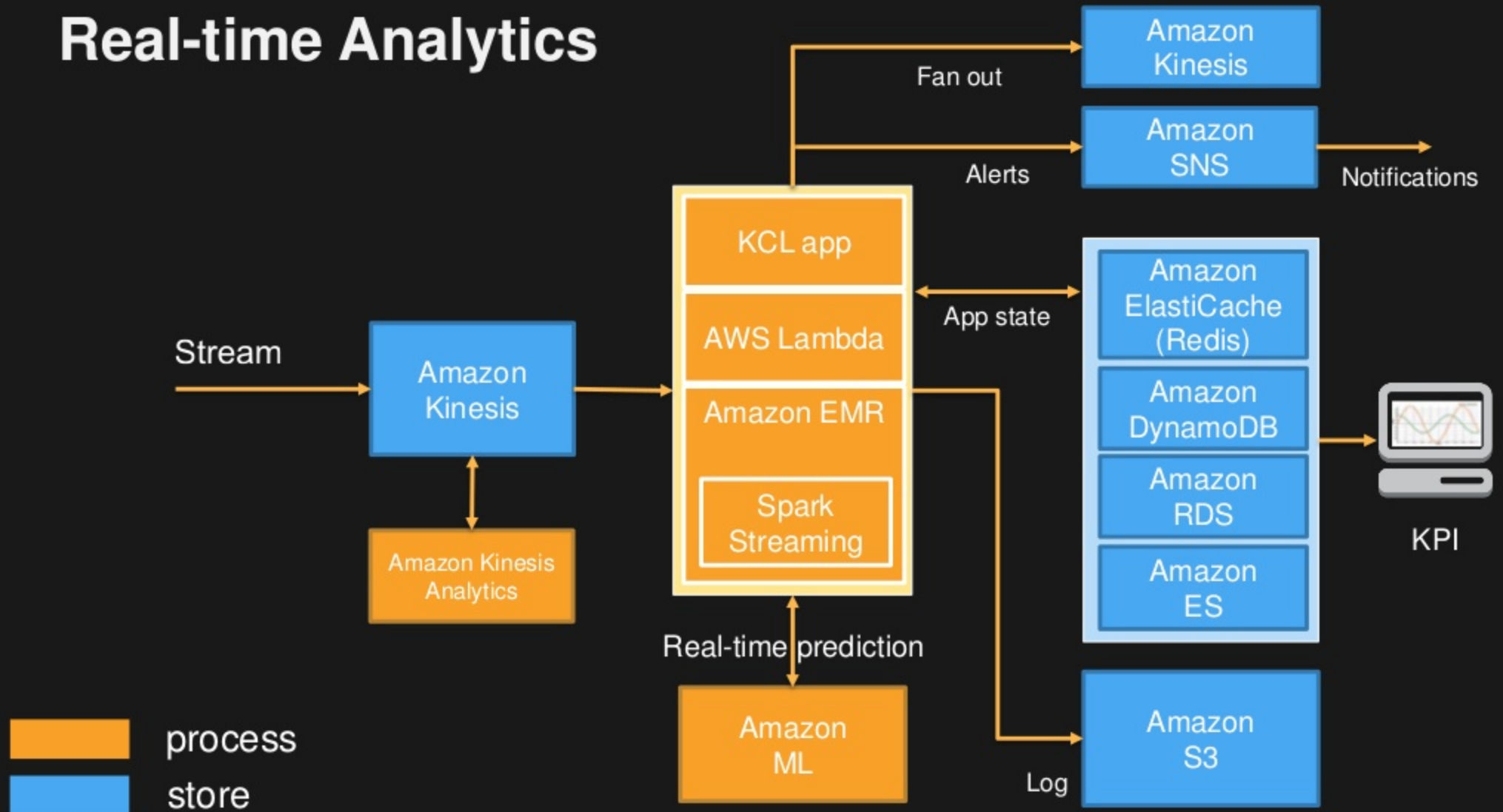
Multiple stages

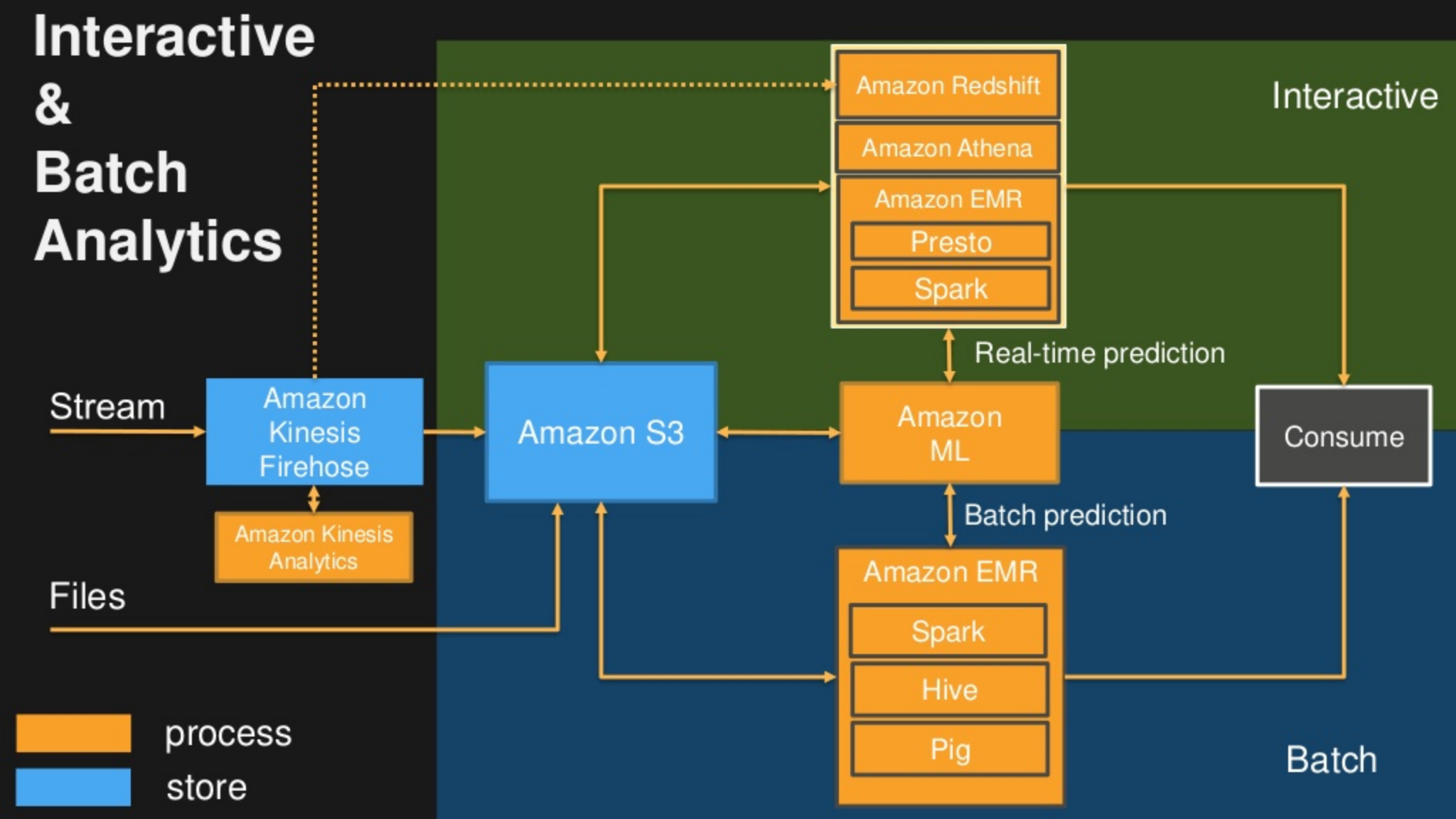
Primitive: Pub/Sub

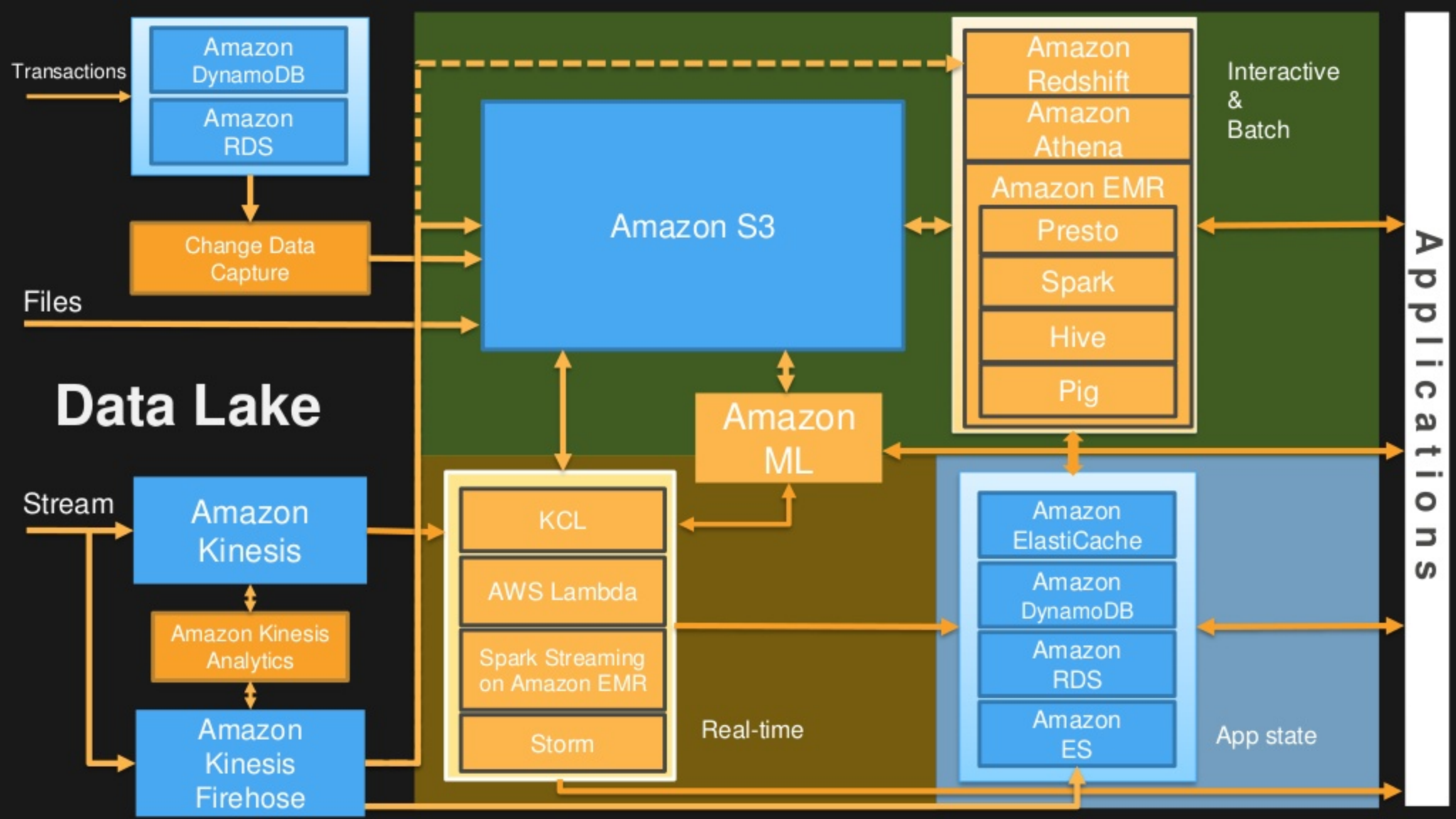
Parallel stream consumption/processing



Real-time Analytics




Interactive & Batch Analytics
Interactive
Amazon Redshift
Amazon Athena
Amazon EMR
Presto
Spark
Real-time prediction
Stream
Amazon Kinesis Firehose
Amazon Kinesis Analytics
Amazon S3
Amazon ML
Consume
Batch prediction
Files
Amazon EMR
Spark
Hive
Pig
Batch
process
store


Transactions
Amazon DynamoDB
Amazon RDS
Change Data Capture
Files
Data Lake
Stream
Amazon Kinesis
Amazon Kinesis Analytics
Amazon Kinesis Firehose
Amazon S3
Amazon ML
KCL
AWS Lambda
Spark Streaming on Amazon EMR
Storm
Real-time
Amazon Redshift
Amazon Athena
Amazon EMR
Presto
Spark
Hive
Pig
Interactive & Batch
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon ES
App state
Applications

Data Lake Solution Architecture on AWS

http://bit.ly/DataLakeOnAWS

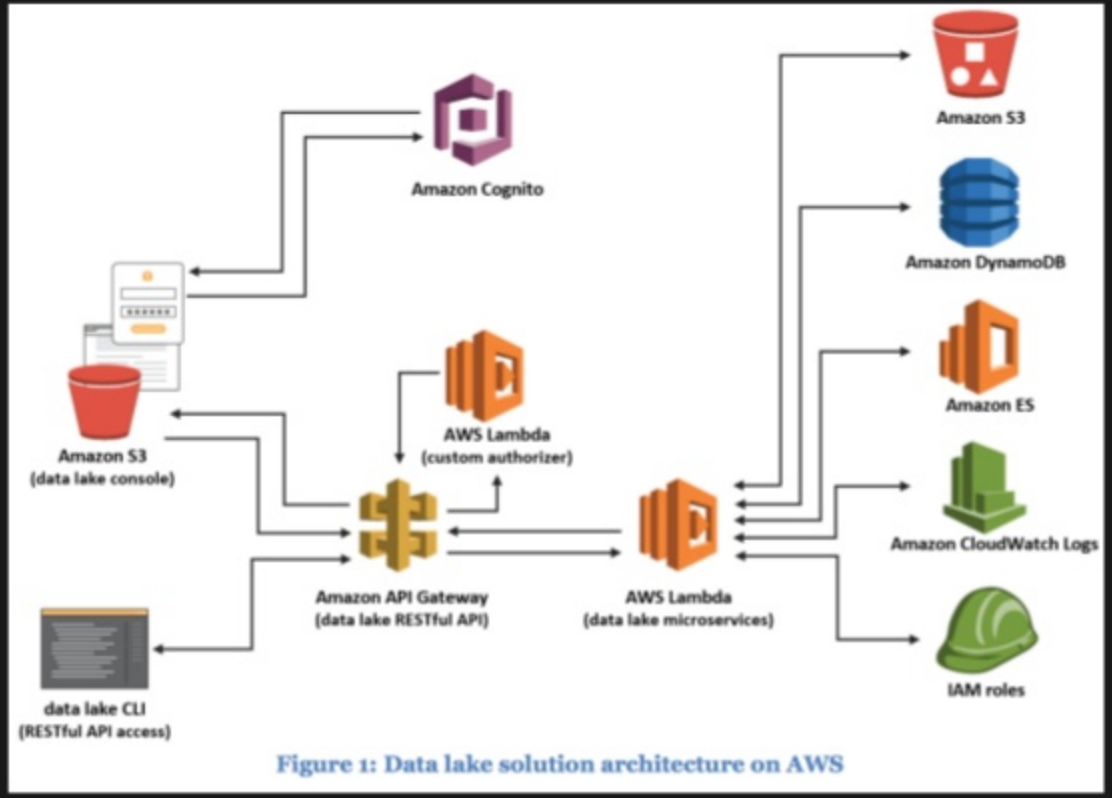
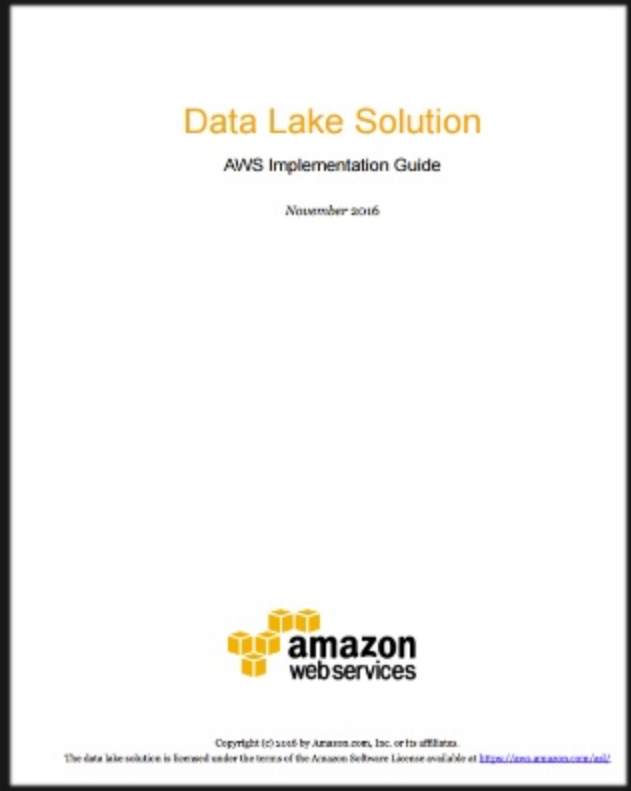


Figure 1: Data lake solution architecture on AWS



AWS

SUMMIT

Thank You

http://blogs.aws.amazon.com/bigdata/

unni_k_pillai

amazon web services