Simplify Big Data with AWS

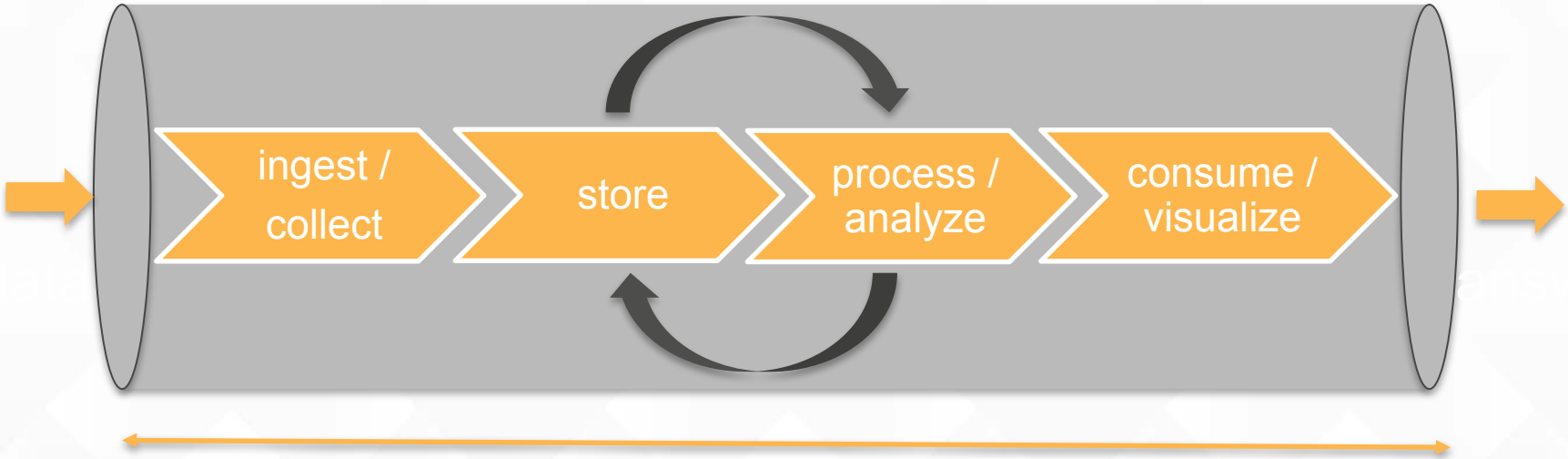
Julien Simon, Principal Technical Evangelist

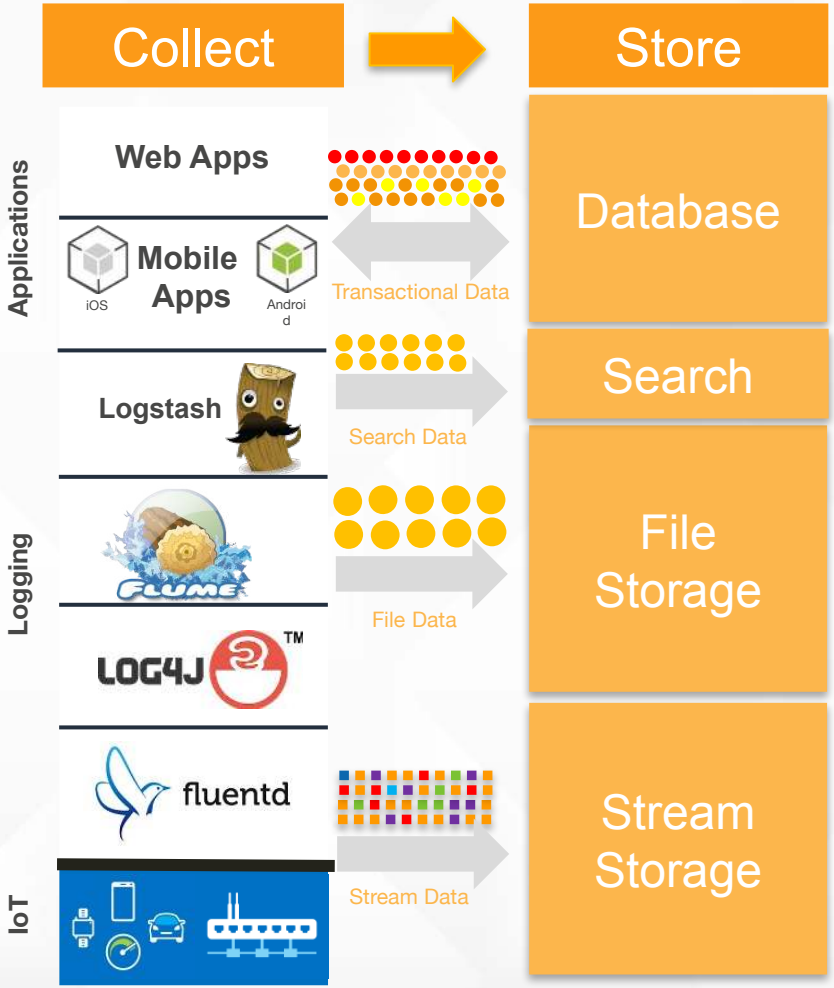
@julsimon

Webinar “Salon du Big Data”
02/03/2016

amazon web services

Simplify Big Data Processing

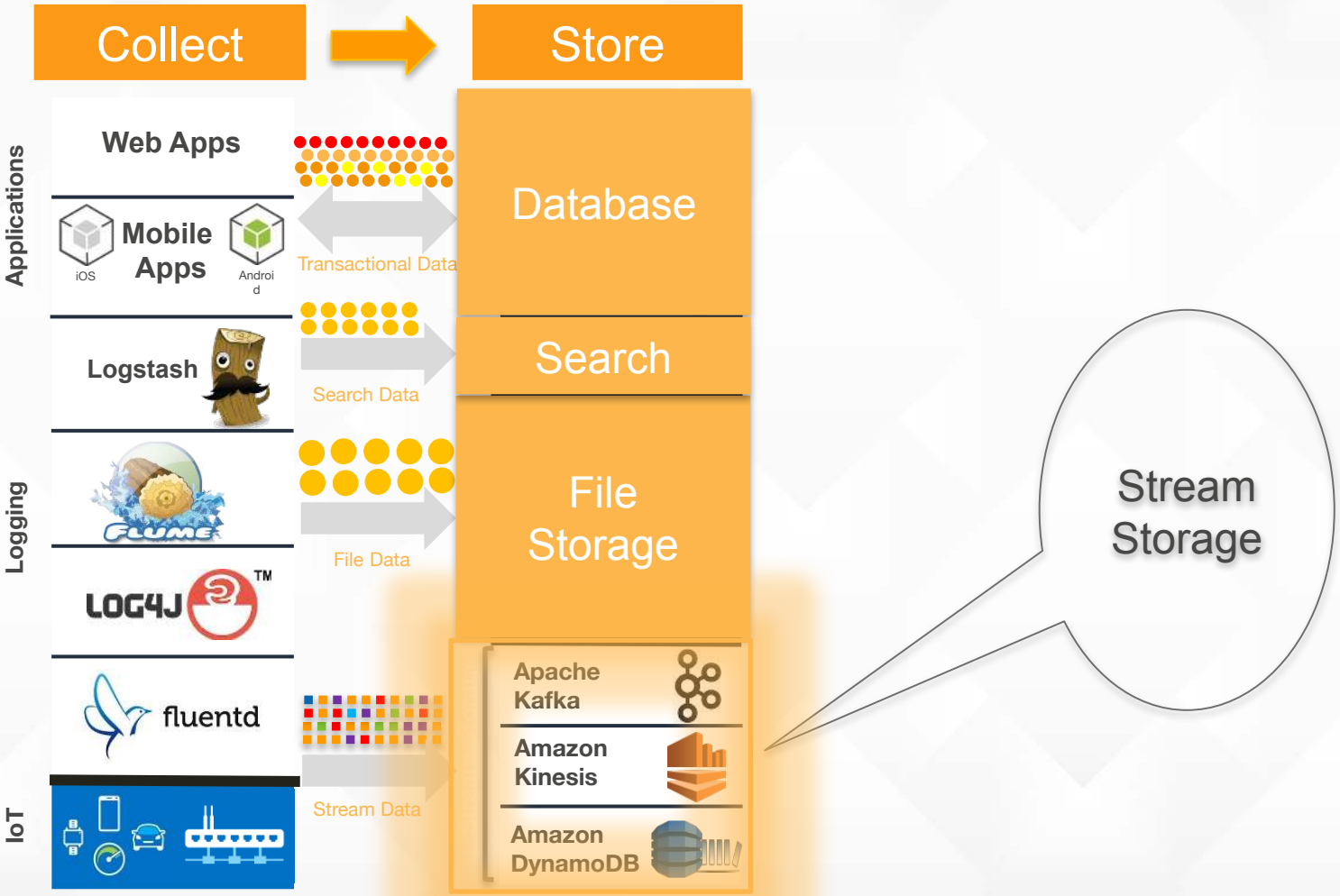




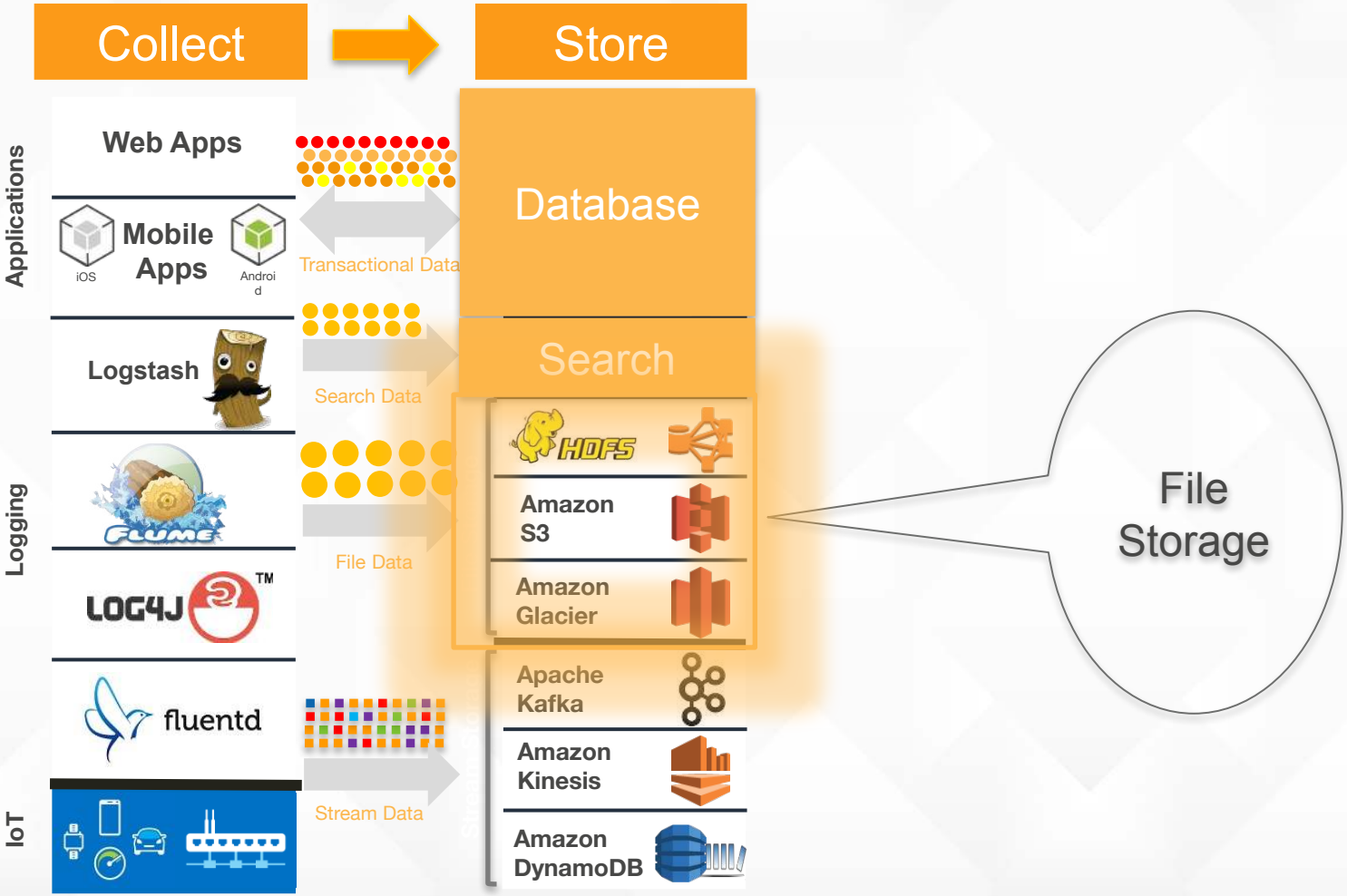
Types of Data

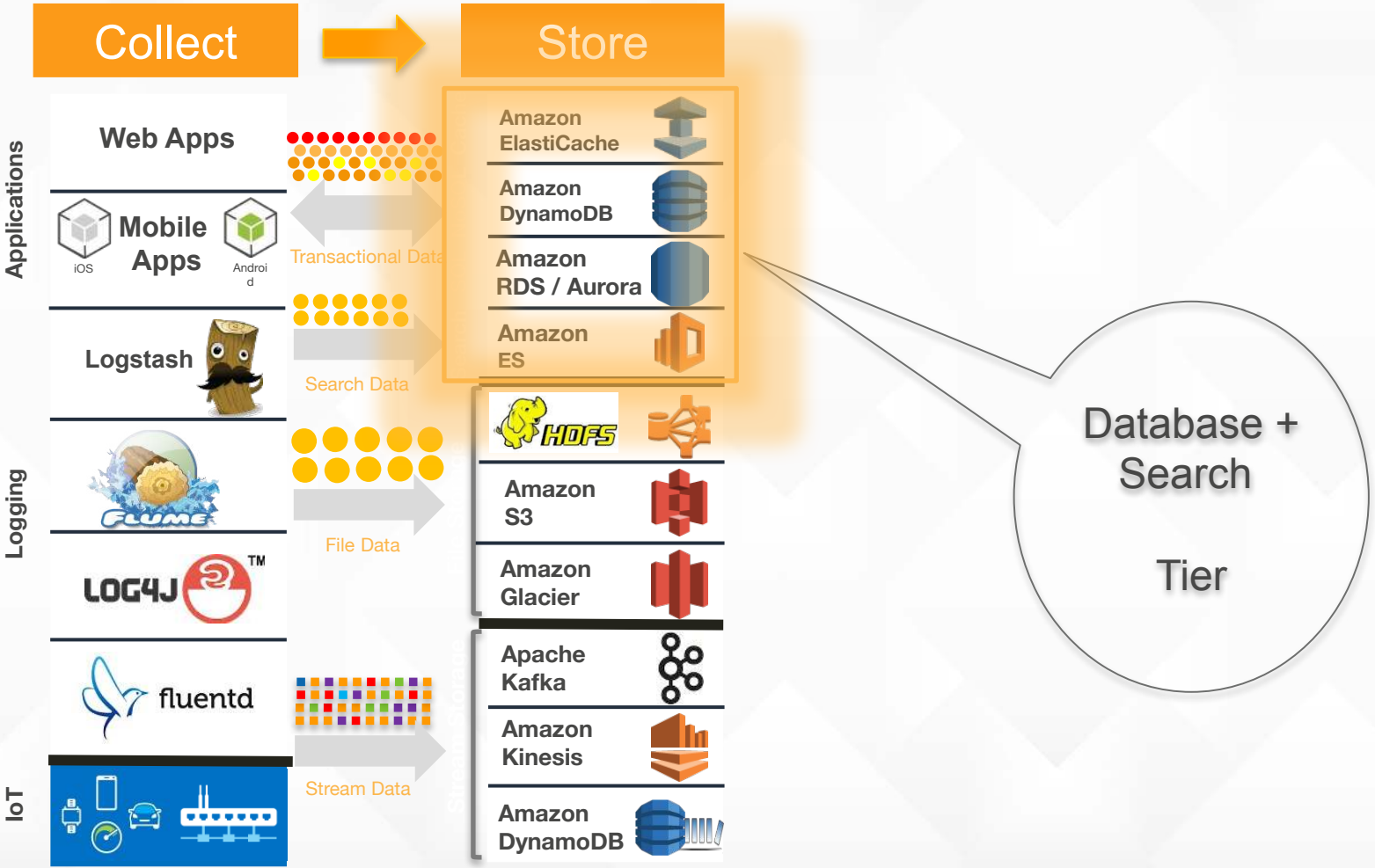
- Transactional
  - Database reads & writes (OLTP)
  - Cache
- Search
  - Logs
  - Streams
- File
  - Log files (/var/log)
  - Log collectors & frameworks
- Stream
  - Log records
  - Sensors & IoT data

amazon web services



amazon web services





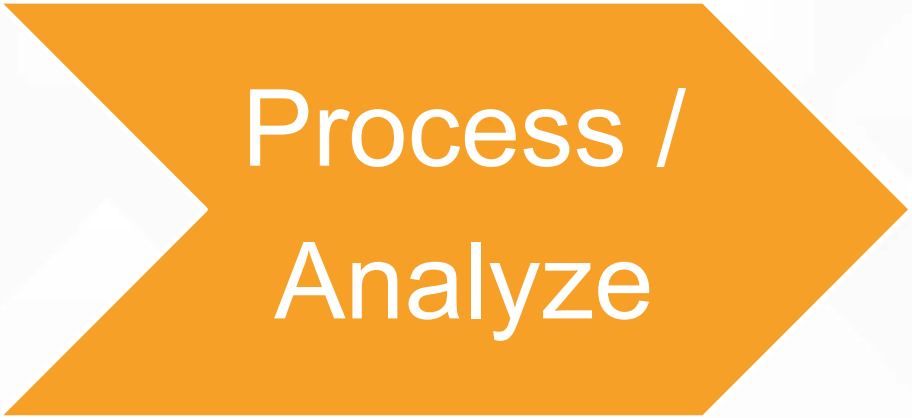
Database + Search Tier Anti-pattern

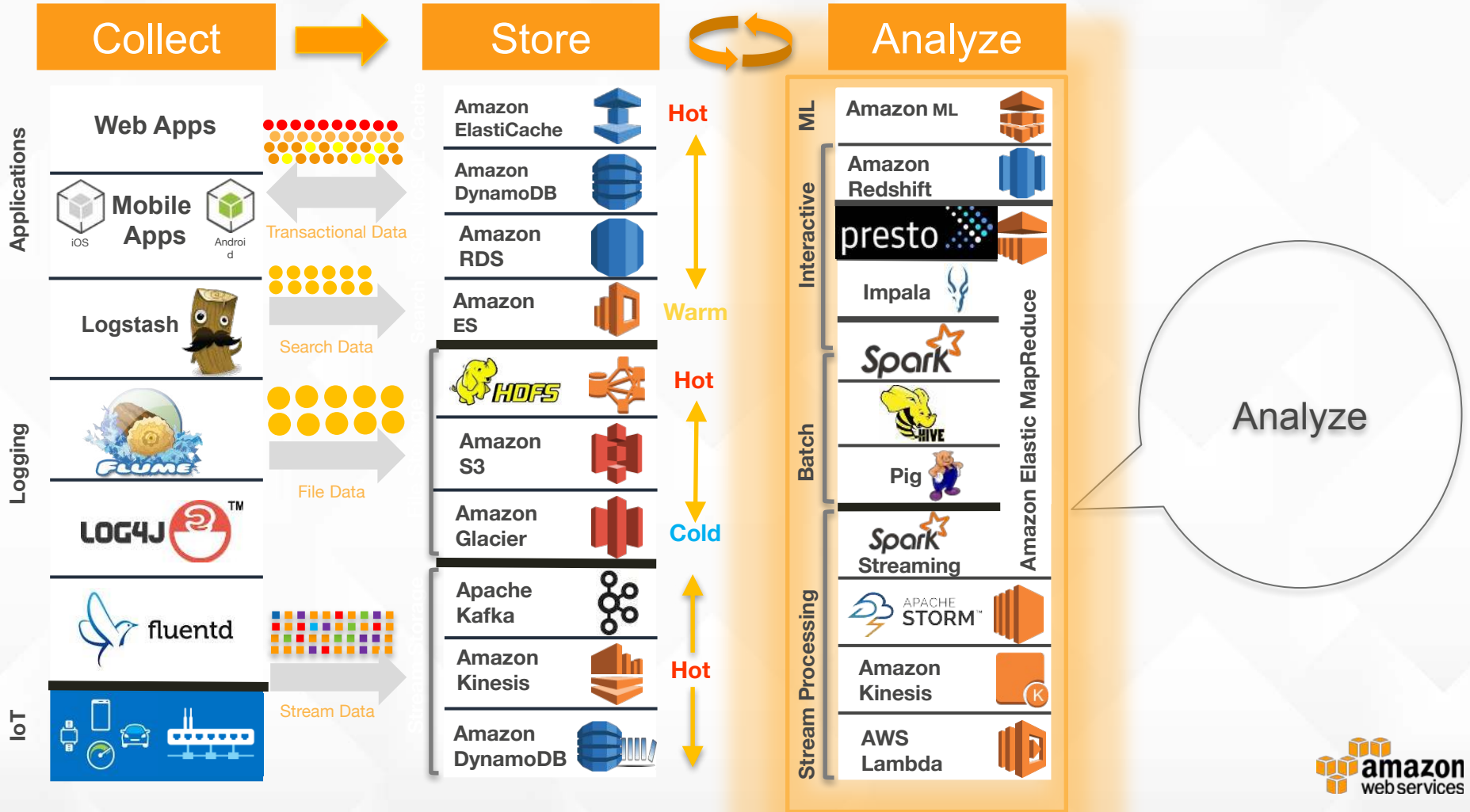
What Data Store Should I Use?

Hot Data — Warm Data — Cold Data

	Amazon ElastiCache	Amazon DynamoDB	Amazon Aurora	Amazon Elasticsearch	Amazon EMR (HDFS)	Amazon S3	Amazon Glacier
Average latency	ms	ms	ms, sec	ms,sec	sec,min,hrs	ms,sec,min (~ size)	hrs
Data volume	GB	GB–TBs (no limit)	GB–TB (64 TB Max)	GB–TB	GB–PB (~nodes)	MB–PB (no limit)	GB–PB (no limit)
Item size	B-KB	KB (400 KB max)	KB (64 KB)	KB (1 MB max)	MB-GB	KB-GB (5 TB max)	GB (40 TB max)
Request rate	High - Very High	Very High (no limit)	High	High	Low – Very High	Low – Very High (no limit)	Very Low
Storage cost GB/month	$$	¢¢	¢¢	¢¢	¢	¢	¢/10
Durability	Low - Moderate	Very High	Very High	High	High	Very High	Very High

Hot Data — Warm Data — Cold Data




Collect
Store
Analyze
Applications
Logging
IoT
Web Apps
Mobile Apps
iOS
Android
Logstash
FLUME
LOG4J
fluentd
Transactional Data
Search Data
File Data
Stream Data
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon ES
HDFS
Amazon S3
Amazon Glacier
Apache Kafka
Amazon Kinesis
Amazon DynamoDB
Hot
Warm
Hot
Cold
Hot
ML
Amazon ML
Interactive
Amazon Redshift
presto
Impala
Batch
Spark
HIVE
Pig
Amazon Elastic MapReduce
Stream Processing
Spark Streaming
APACHE STORM
Amazon Kinesis
AWS Lambda
Analyze
amazon web services

Analysis Tools and Frameworks

Machine Learning

- Mahout, Spark ML, Amazon ML

Interactive Analytics

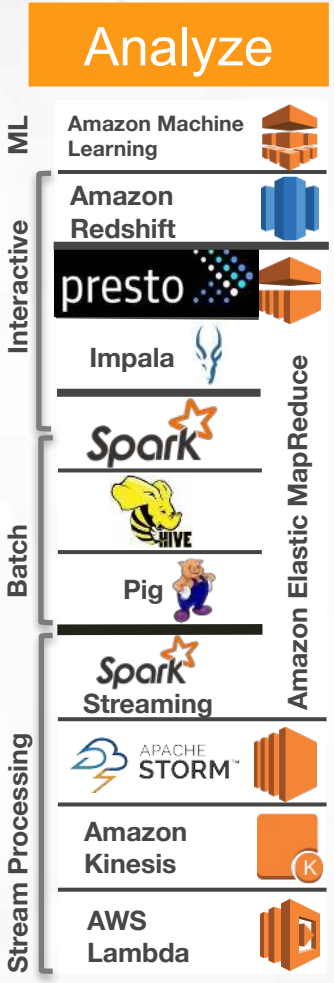
- Amazon Redshift, Presto, Impala, Spark

Batch Processing

- MapReduce, Hive, Pig, Spark

Stream Processing

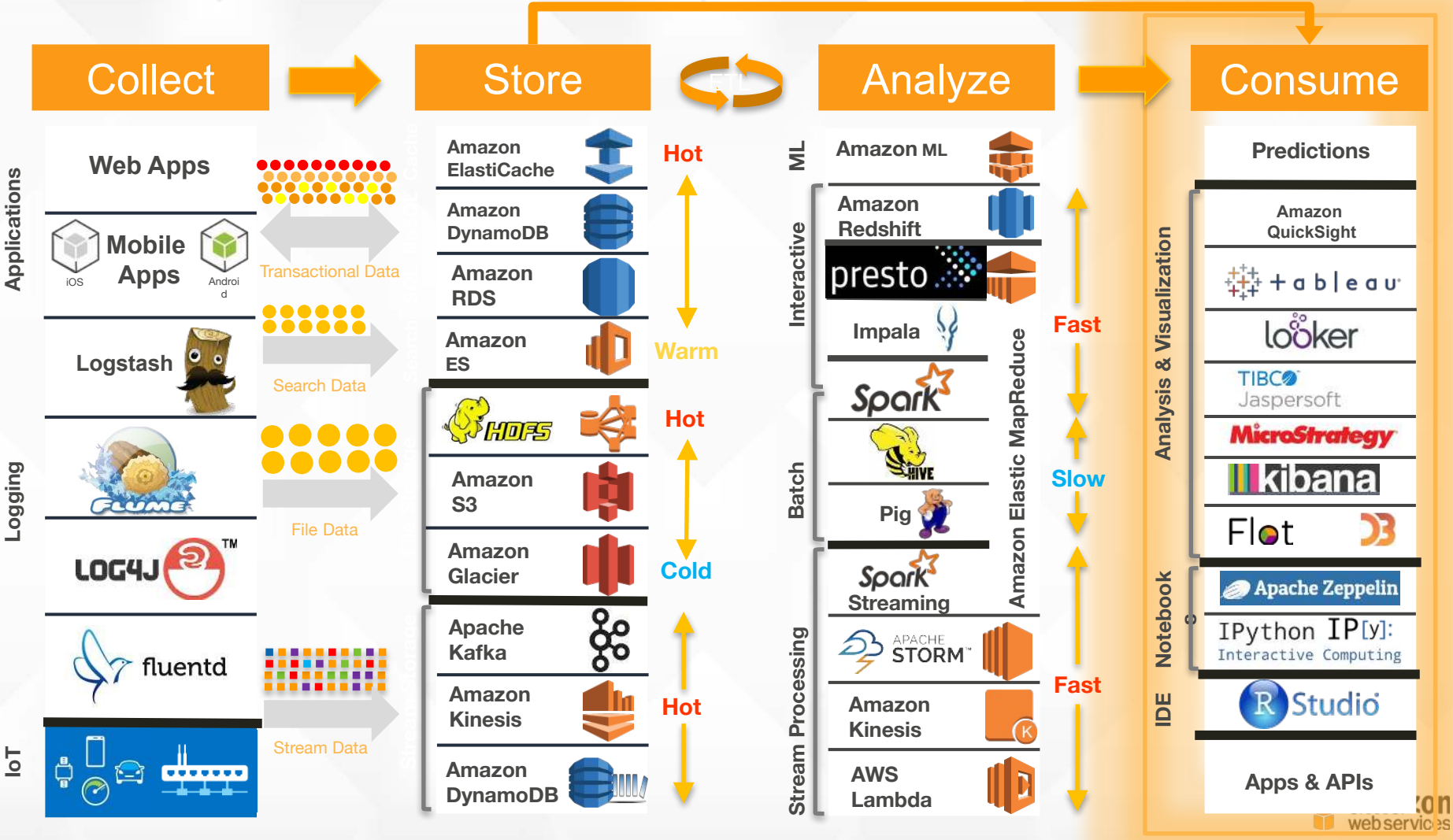
- Micro-batch: Spark Streaming, KCL, Hive, Pig
- Real-time: Storm, AWS Lambda, KCL



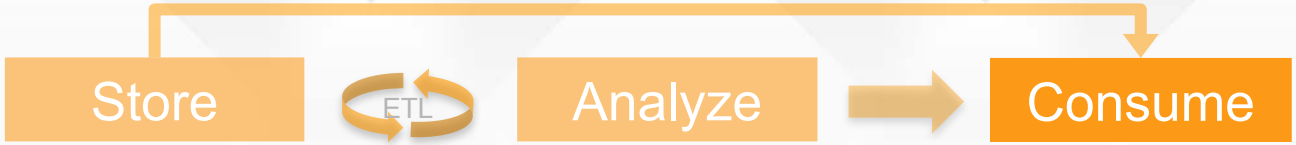
What Data Processing Technology Should I Use?

	Amazon Redshift	Impala	Presto	Spark	Hive
Query Latency	Low	Low	Low	Low	Medium (Tez) – High (MapReduce)
Durability	High	High	High	High	High
Data Volume	1.6 PB Max	~Nodes	~Nodes	~Nodes	~Nodes
Managed	Yes	Yes (EMR)	Yes (EMR)	Yes (EMR)	Yes (EMR)
Storage	Native	HDFS / S3	HDFS / S3	HDFS / S3	HDFS / S3
SQL Compatibility	High	Medium	High	Low (SparkSQL)	Medium (HQL)

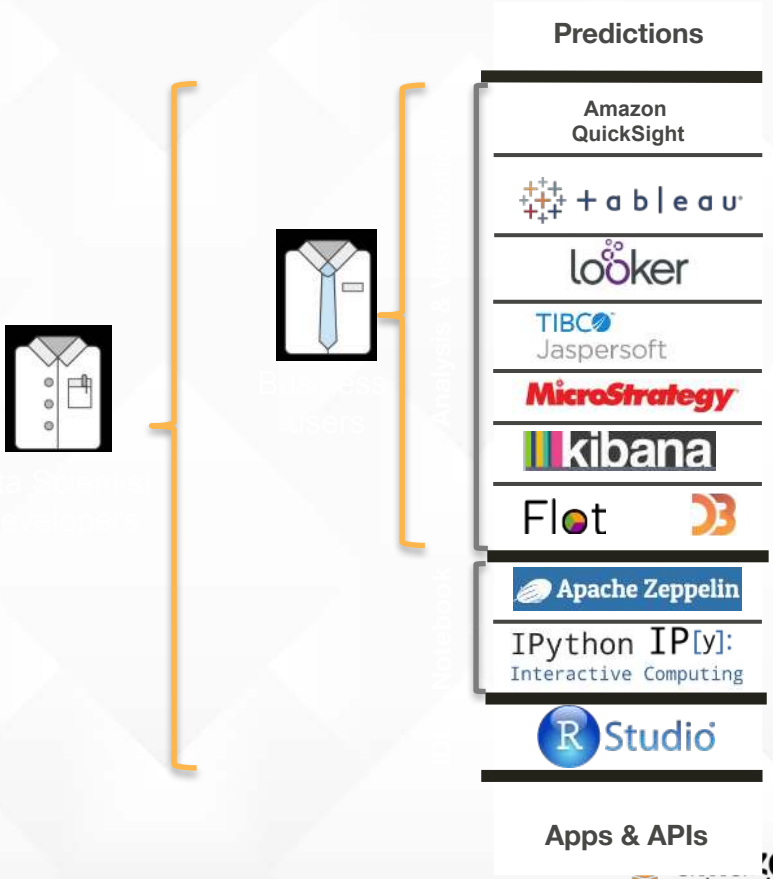
Query Latency (Low is better)


Collect
Store
ETL
Analyze
Consume
Applications
Web Apps
Mobile Apps
iOS
Android
Logging
Logstash
FLUME
LOG4J
fluentd
IoT
Transactional Data
Search Data
File Data
Stream Data
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon ES
HDFS
Amazon S3
Amazon Glacier
Apache Kafka
Amazon Kinesis
Amazon DynamoDB
Hot
Warm
Hot
Cold
Hot
ML
Amazon ML
Interactive
Amazon Redshift
presto
Impala
Batch
Spark
HIVE
Pig
Stream Processing
Spark Streaming
APACHE STORM
Amazon Kinesis
AWS Lambda
Amazon Elastic MapReduce
Fast
Slow
Fast
Predictions
Analysis & Visualization
Amazon QuickSight
+ableau
looker
TIBCO Jaspersoft
MicroStrategy
kibana
Flot
IDE
Notebook
Apache Zeppelin
IPython IP[y]: Interactive Computing
RStudio
Apps & APIs
web services

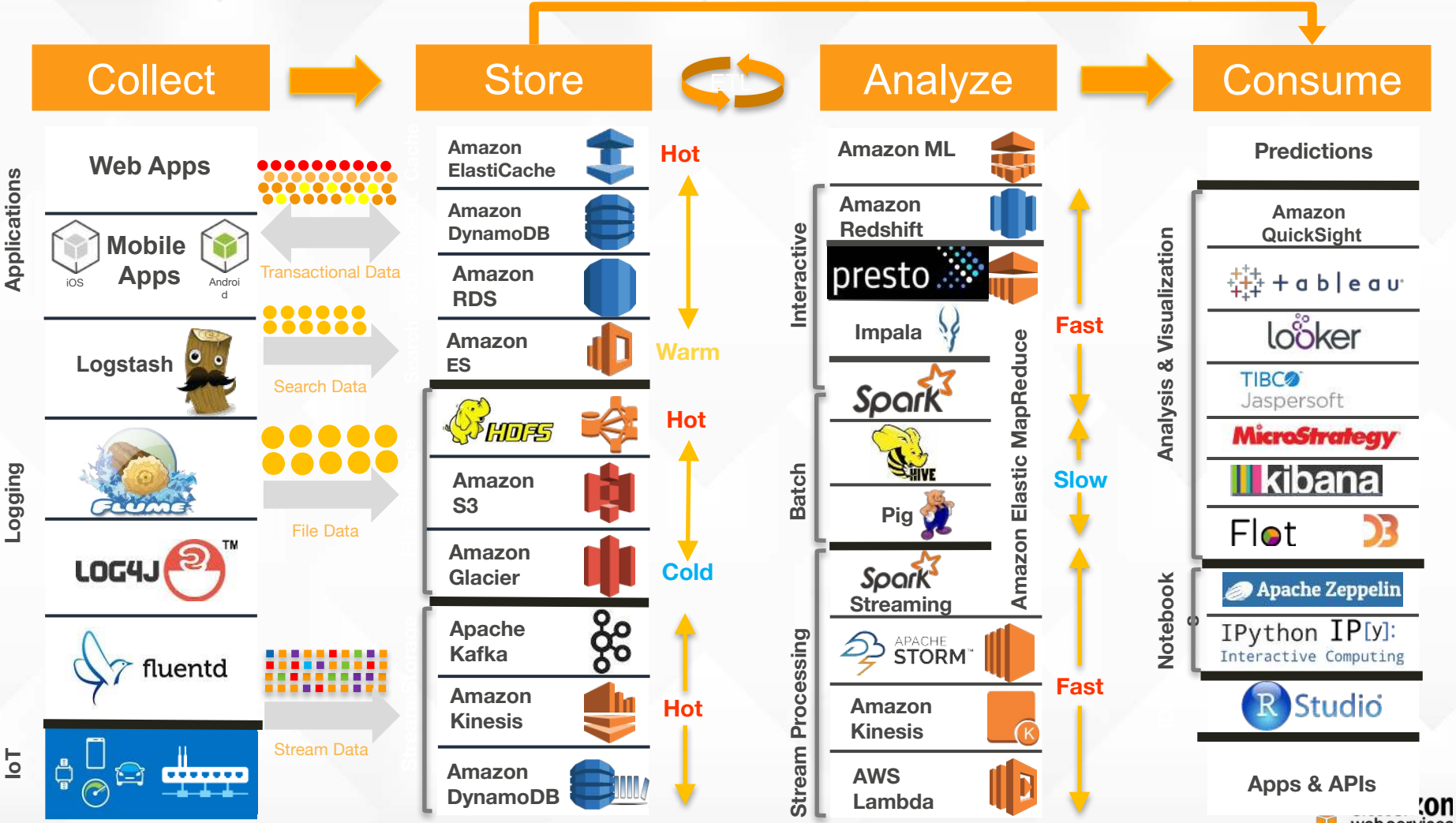
Consume



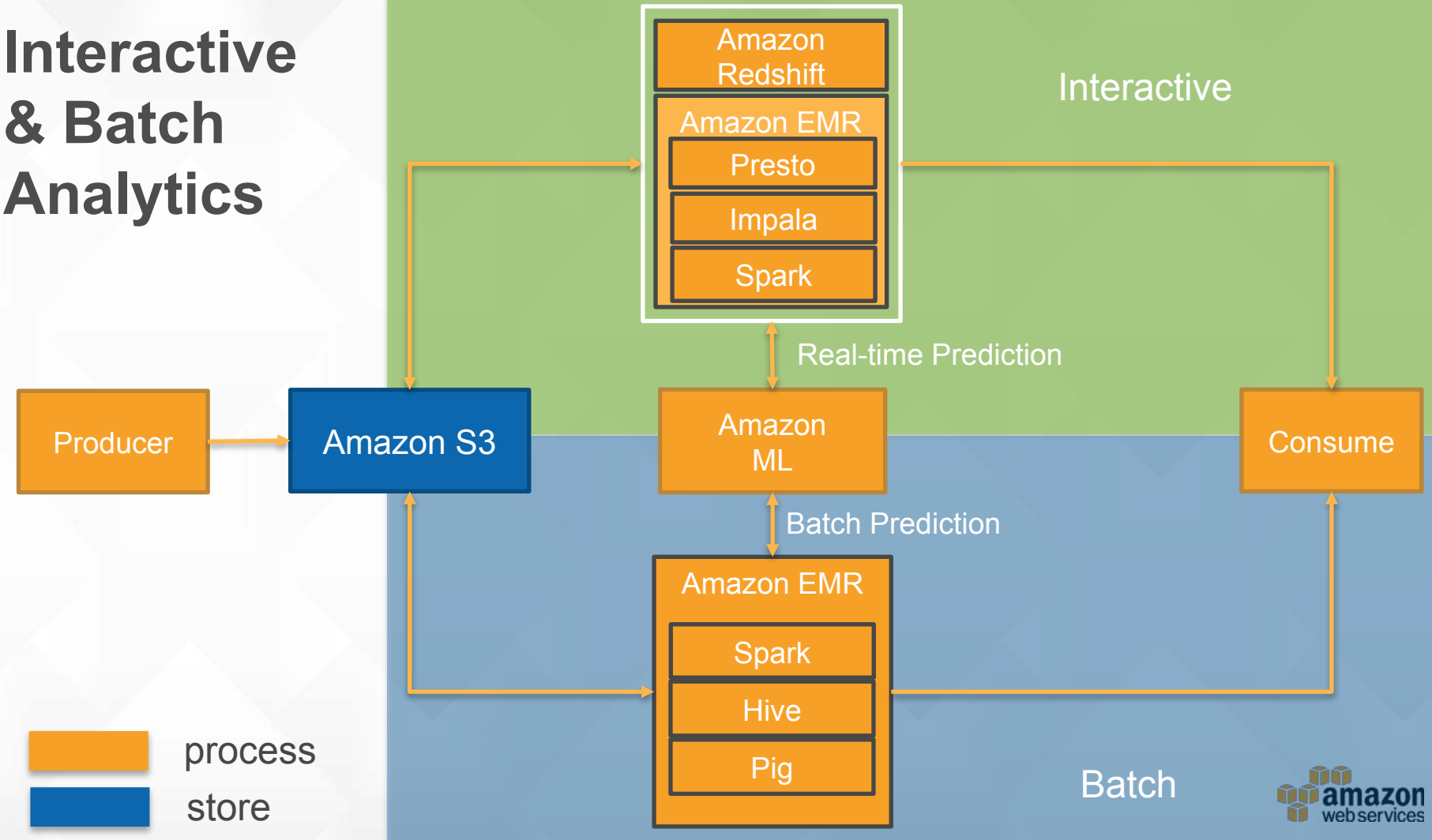
- Predictions
- Analysis and Visualization
- Notebooks
- IDE
- Applications & API



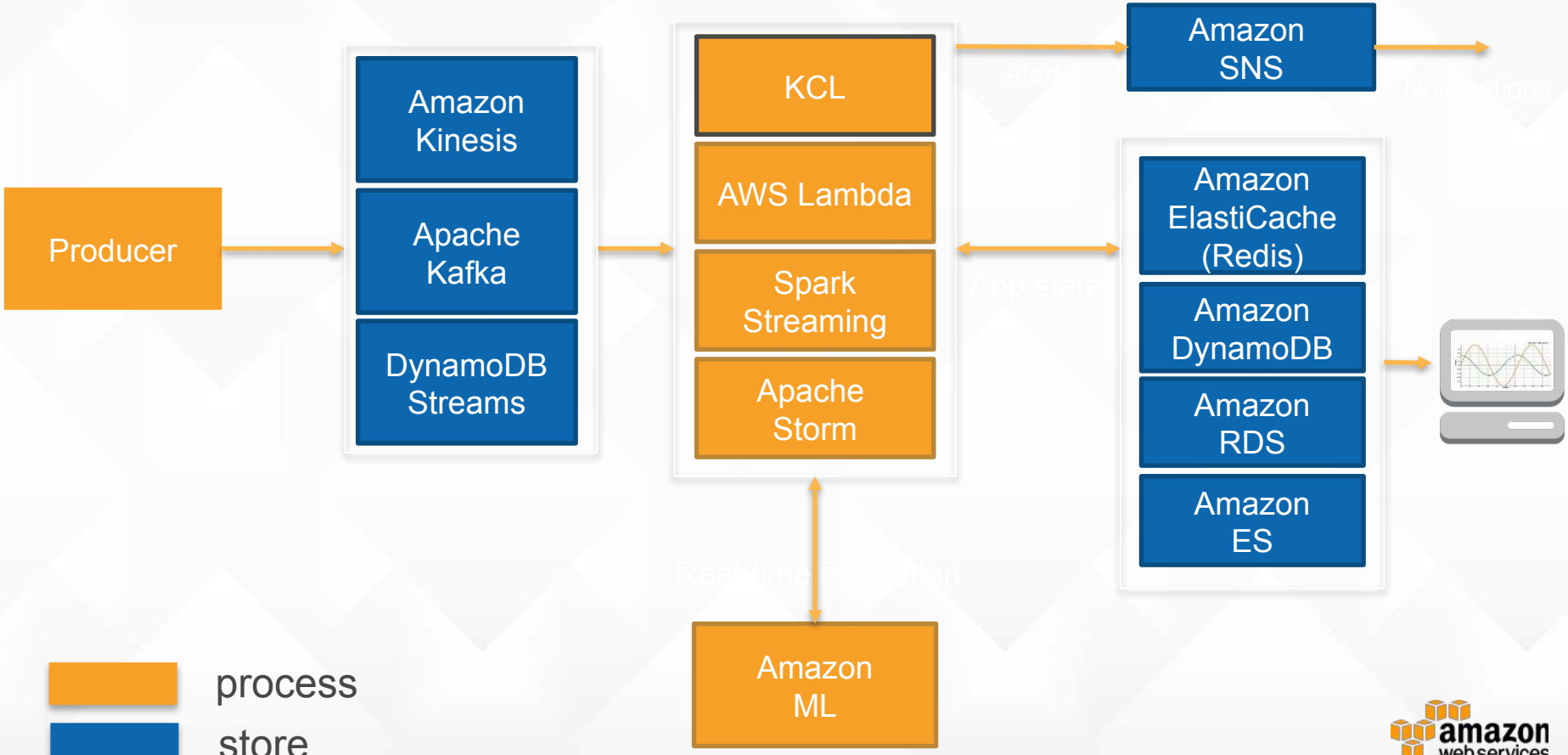
Putting It All Together


Collect
Store
ETL
Analyze
Consume
Applications
Logging
IoT
Web Apps
Mobile Apps
iOS
Android
Logstash
fluentd
Transactional Data
Search Data
File Data
Stream Data
Amazon ElastiCache
Amazon DynamoDB
Amazon RDS
Amazon ES
HDFS
Amazon S3
Amazon Glacier
Apache Kafka
Amazon Kinesis
Amazon DynamoDB
Hot
Warm
Hot
Cold
Hot
Amazon ML
Amazon Redshift
presto
Impala
Spark
HIVE
Pig
Spark Streaming
APACHE STORM
Amazon Kinesis
AWS Lambda
Interactive
Batch
Stream Processing
Amazon Elastic MapReduce
Fast
Slow
Fast
Predictions
Amazon QuickSight
tableau
looker
TIBCO Jaspersoft
MicroStrategy
kibana
Flot
Apache Zeppelin
IPython IP[y]: Interactive Computing
RStudio
Apps & APIs
Analysis & Visualization
Notebook
web services

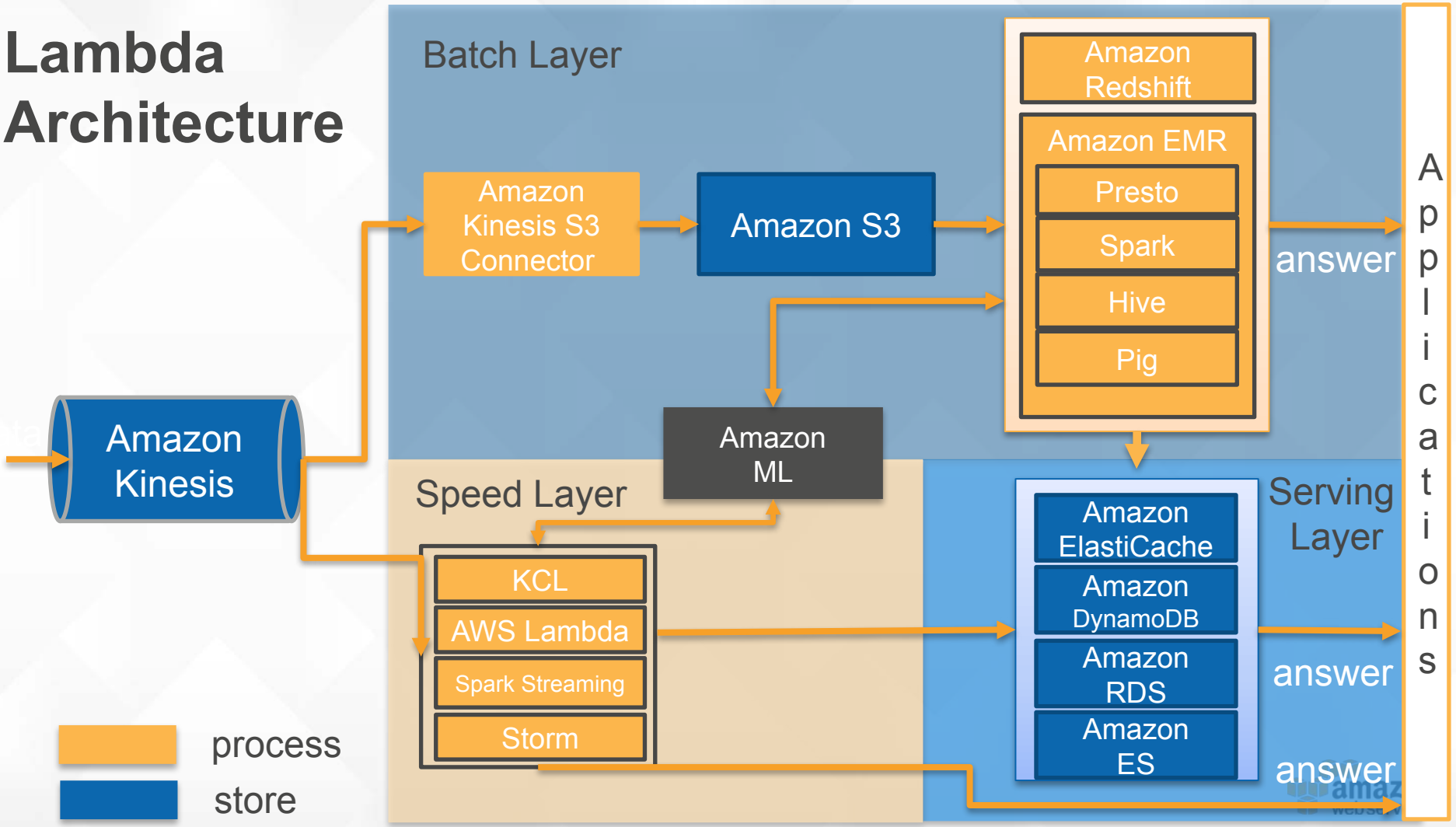
Interactive & Batch Analytics



Real-time Analytics



Lambda Architecture



Summary

- Use the right tool for the job
  - Latency, throughput, access patterns
- Leverage AWS managed services
  - No/low admin
- Be cost conscious
  - Big data ≠ big cost

Thank you. Let's keep in touch!

@aws_actus @julsimon
facebook.com/groups/AWSFrance/

AWS User Groups in Paris, Lyon, Nantes, Lille & Rennes (meetup.com)

March 7-8

AWSome Day

March 16

March 23-24

April 6-7 (Lyon)

April 20-22

April 25

AWS Summit
May 31st

amazon web services

Customer references & further reading

- Amazon Kinesis: https://aws.amazon.com/solutions/case-studies/supercell/
- Amazon DynamoDB: https://aws.amazon.com/fr/solutions/case-studies/adroll/
- Amazon S3 / Glacier: https://aws.amazon.com/fr/solutions/case-studies/soundcloud/
- Amazon EMR: https://aws.amazon.com/fr/solutions/case-studies/yelp/
- Amazon Aurora: https://aws.amazon.com/fr/rds/aurora/testimonials/
- Amazon Redshift: https://aws.amazon.com/fr/solutions/case-studies/financial-times/
- AWS Lambda: https://aws.amazon.com/fr/solutions/case-studies/nordstrom/
- Many more case studies at https://aws.amazon.com/fr/solutions/case-studies/big-data/

- Whitepaper: “Big Data Analytics Options on AWS” : http://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf
- AWS Big Data blog: https://blogs.aws.amazon.com/bigdata