AWS INNOVATE

Databases and Analytics on the AWS Cloud

Antoine Généreux, AWS Solutions Architect

May 10, 2017

What to expect from the session

AWS INNOVATE

- Challenges and architectural principles
- How to simplify big data processing
- What database technologies should you use?
  - Why and When?
  - How?
- Architectural patterns and Customer Examples

amazon web services | intel

Key Ideas

AWS INNOVATE

- Data is your organization's most valuable resource
- All data has the potential to be big data
- Databases are no longer the center of analytics*

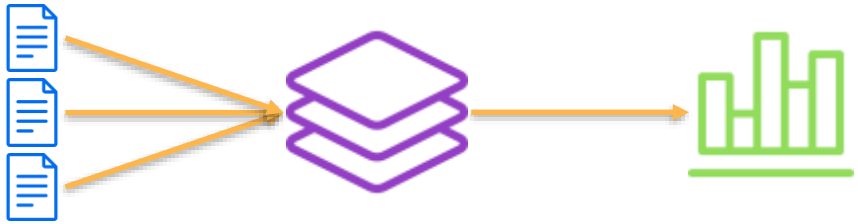
*but they play a critical role!

amazon web services | intel

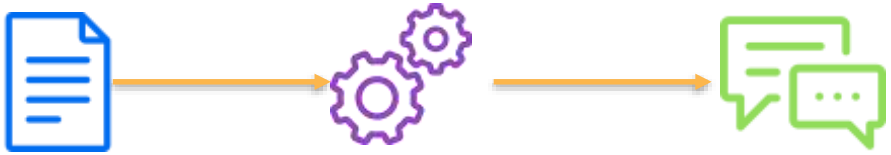
Evolution of Analytics

AWS INNOVATE

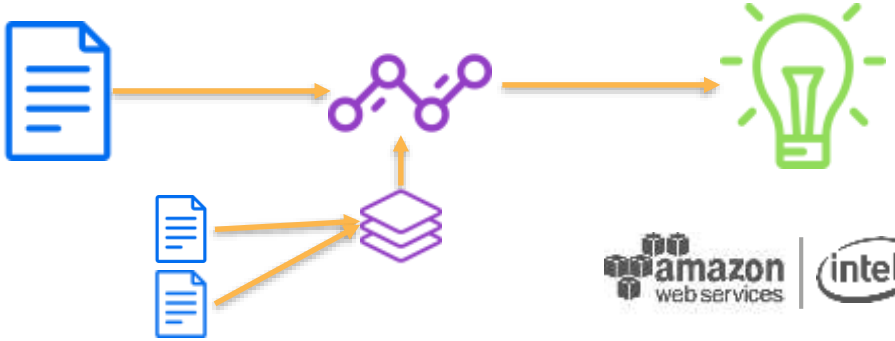
Batch analytics



Real-time analytics



Predictive/Adaptive analytics



amazon web services | intel

A plethora of tools

AWS INNOVATE

Challenges

AWS INNOVATE

- Is there a reference architecture?
- What tools should I use?
- How?
- Why?

amazon web services | intel

Architectural Principles

AWS INNOVATE

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 (schema-on-read)

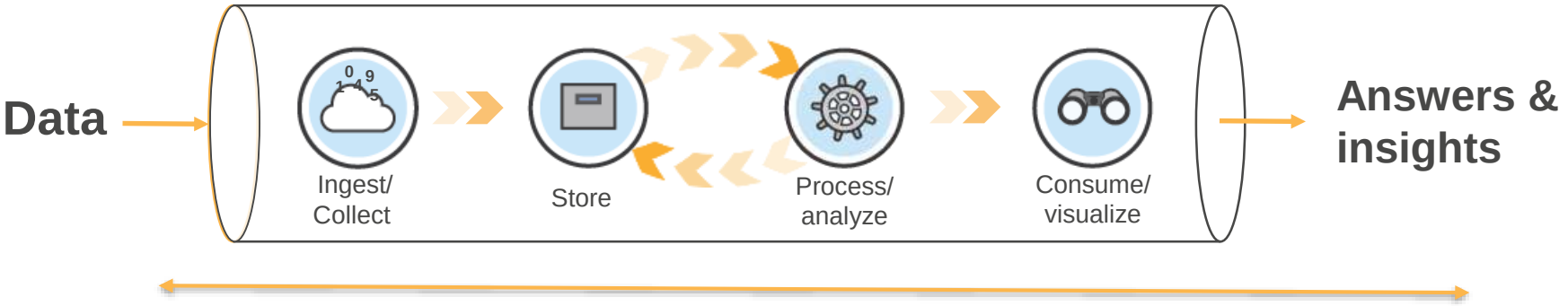
Be cost-conscious

- Big data ≠ big cost

amazon web services | intel

Simplify Big Data Processing

AWS INNOVATE



Time to answer (Latency)

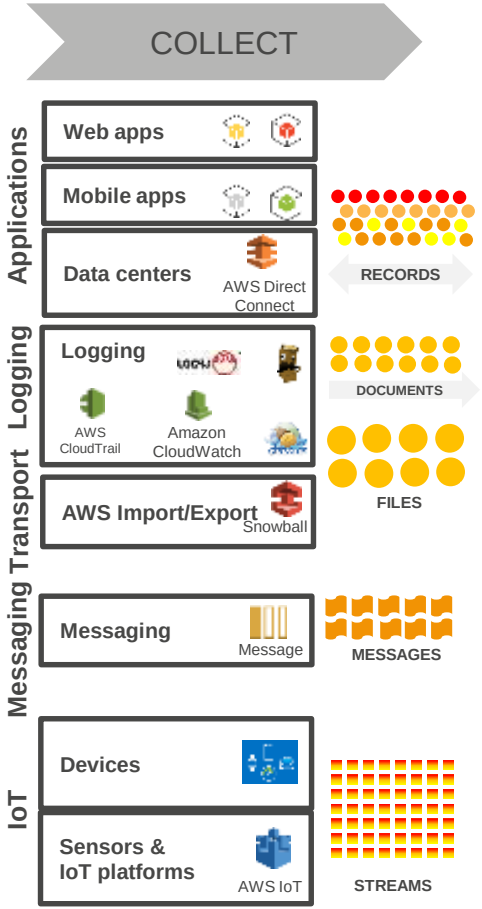
Throughput

Cost

amazon web services | intel

Types of Data

AWS INNOVATE



In-memory data structures

Database records

Transactions

Search documents

Log files

Files

Messages

Data streams

Events

amazon web services | intel

Data Temperature

AWS INNOVATE

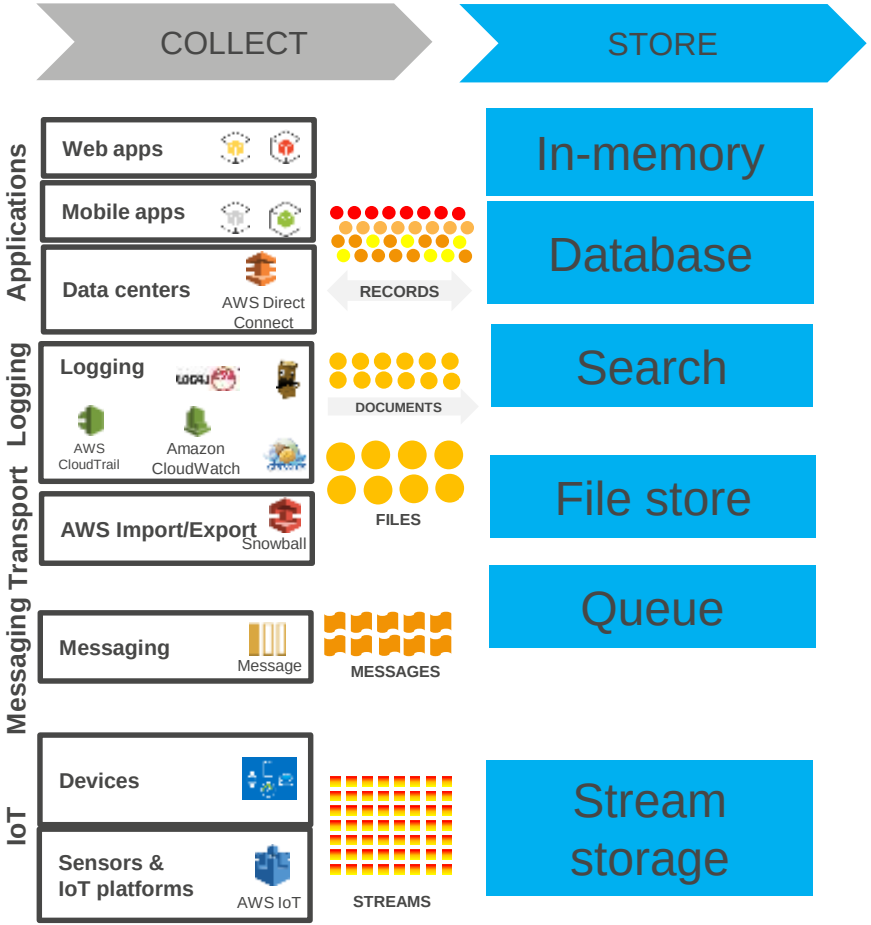
	Hot	Warm	Cold
Volume	MB–GB	GB–TB	PB–EB
Item size	B–KB	KB–MB	KB–TB
Latency	ms	ms, sec	min, hrs
Durability	Low–high	High	Very high
Request rate	Very high	High	Low
Cost/GB	$$-$	$-¢¢	¢

Hot data — Warm data — Cold data

amazon web services | intel

Types of Data Stores

AWS INNOVATE



Caches, data structure servers

SQL & NoSQL databases

Search engines

File systems

Message queues

Pub/sub message queues

amazon web services | intel

Message & Stream Storage

Amazon SQS

- Managed message queue service

Apache Kafka

- High throughput distributed streaming platform

Amazon Kinesis Streams

- Managed stream storage + processing

Amazon Kinesis Firehose

- Managed data delivery

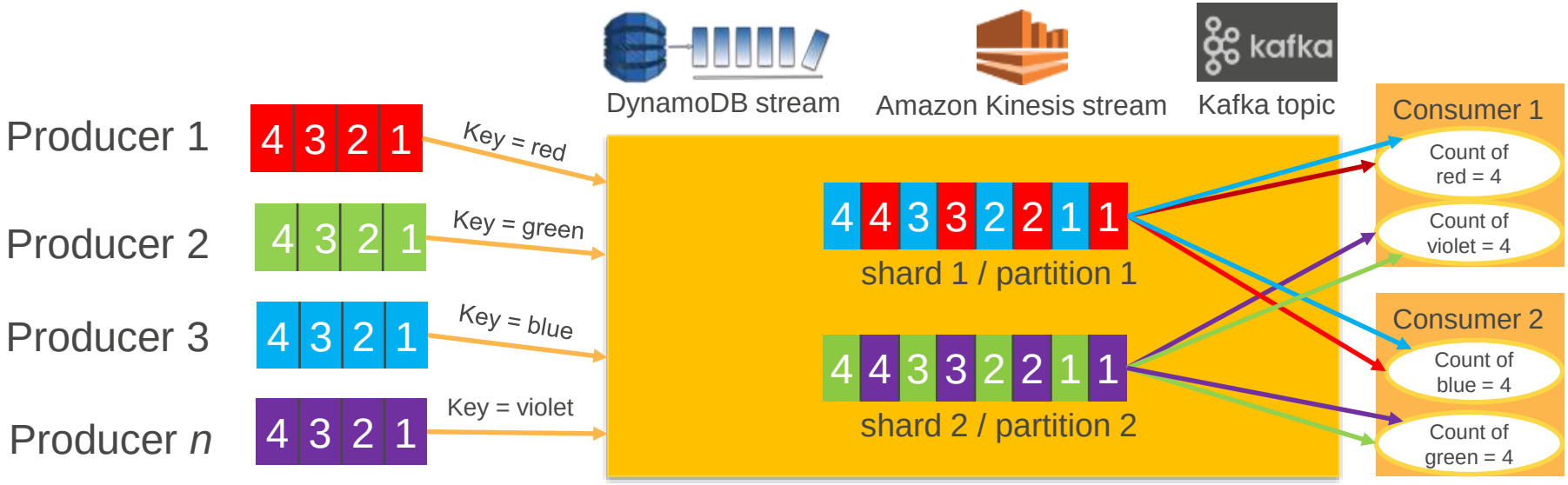
Amazon DynamoDB

- Managed NoSQL database
- Tables can be stream-enabled

Why Stream Storage?

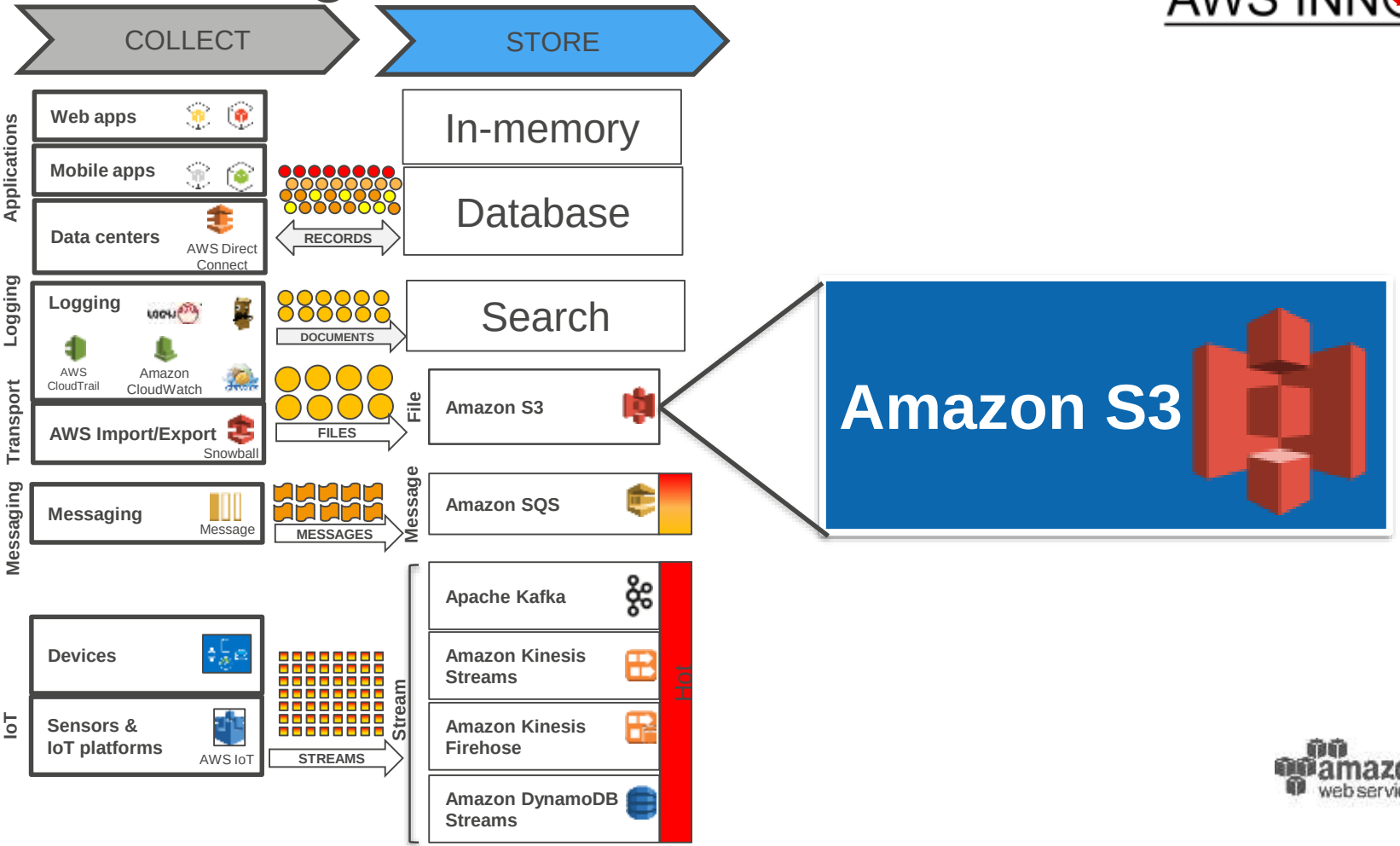
AWS INNOVATE

- Decouple producers & consumers
- Persistent buffer
- Collect multiple streams
- Preserve client ordering
- Parallel consumption
- Streaming MapReduce



amazon web services | intel

File Storage



AWS INNOVATE

amazon web services | intel

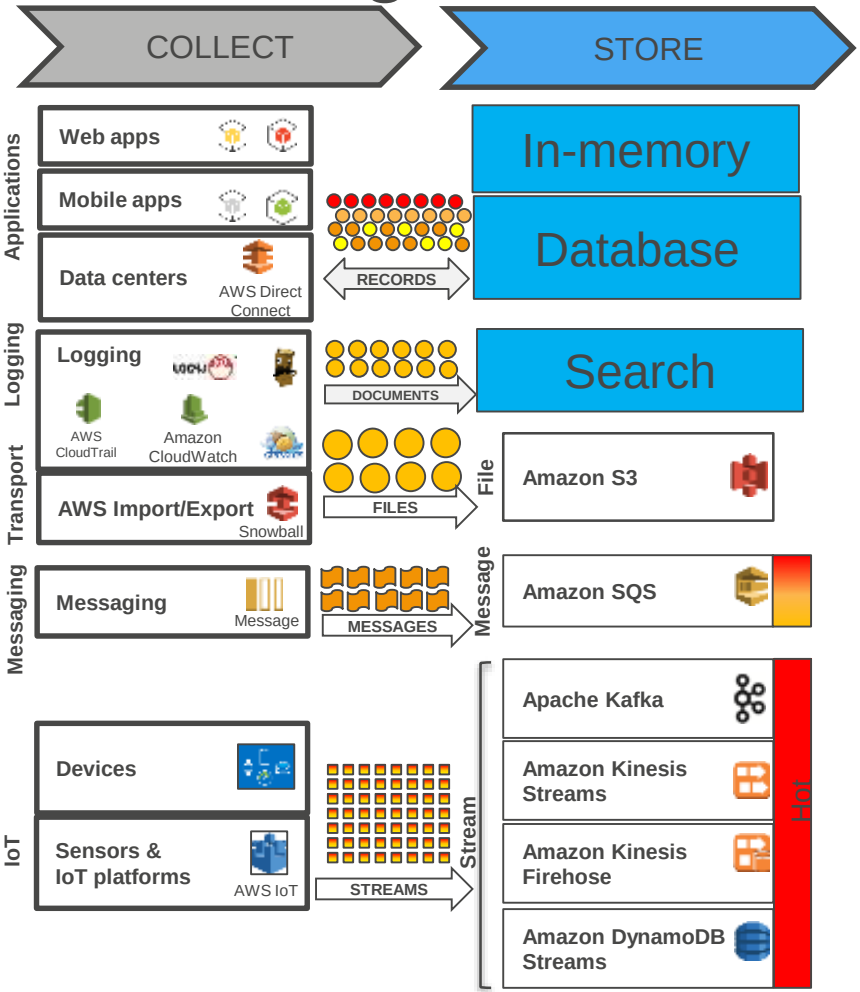
Why Is Amazon S3 Good for Analytics?

AWS INNOVATE

- Natively supported by big data frameworks (Spark, Hive, Presto, etc.)
- No need to run compute clusters for storage (unlike HDFS)
- Can run transient Hadoop clusters & Amazon EC2 Spot Instances
- Multiple & heterogeneous analysis clusters can use the same data
- Unlimited number of objects and volume of data
- Very high bandwidth – no aggregate throughput limit
- 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

amazon web services | intel

File Storage

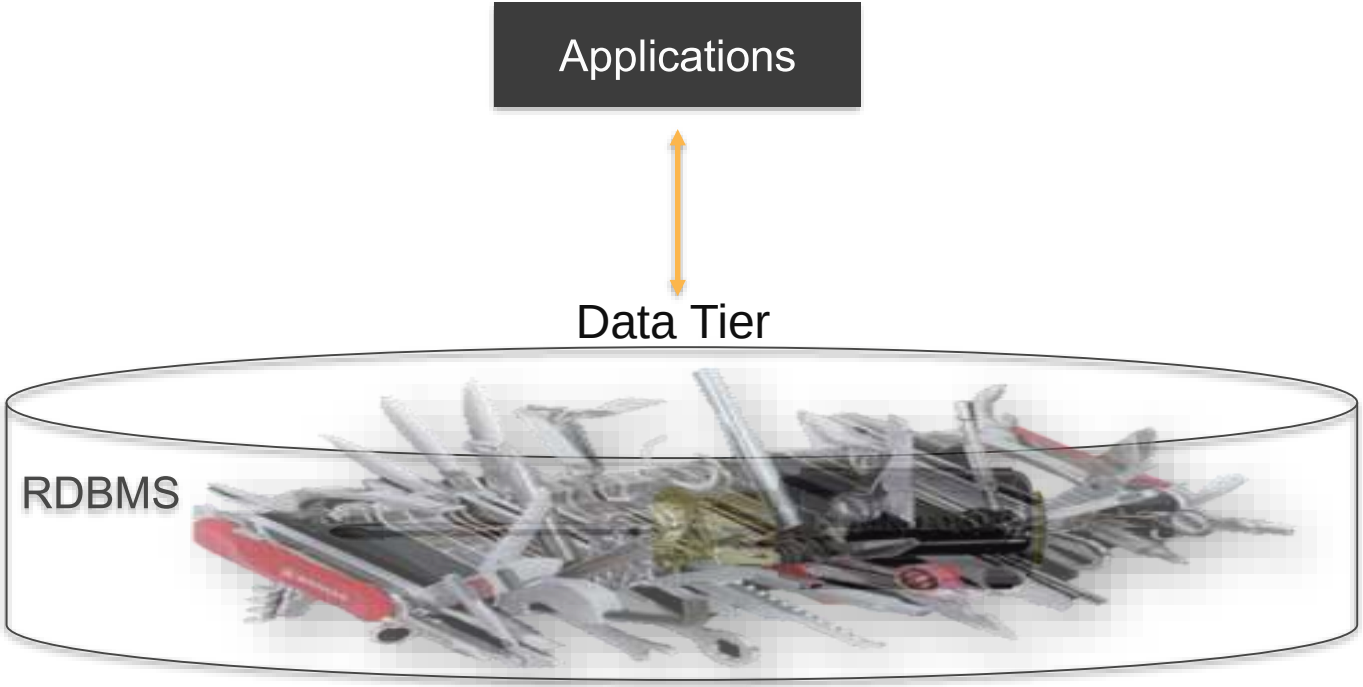


AWS INNOVATE

amazon web services | intel

Database Access Anti-Pattern

AWS INNOVATE



amazon web services | intel

Best Practice: Use the Right Tool for the Job

AWS INNOVATE

Applications

In-memory	NoSQL	SQL	Search
Amazon ElastiCache Redis Memcached	Amazon DynamoDB Cassandra HBase MongoDB	Amazon Aurora Amazon RDS MySQL PostgreSQL Oracle SQL Server	Amazon Elasticsearch Service

amazon web services | intel

Amazon ElastiCache

Microsecond Real-Time Performance

Fully Managed

Redis Automatic Failover = NoOps

Enhanced Redis Engine

No Cross-AZ Data Transfer Costs

Easy to Deploy, Use and Monitor

Open-Source Compatible

Amazon DynamoDB

Fully managed NoSQL

Document / Key-Value store

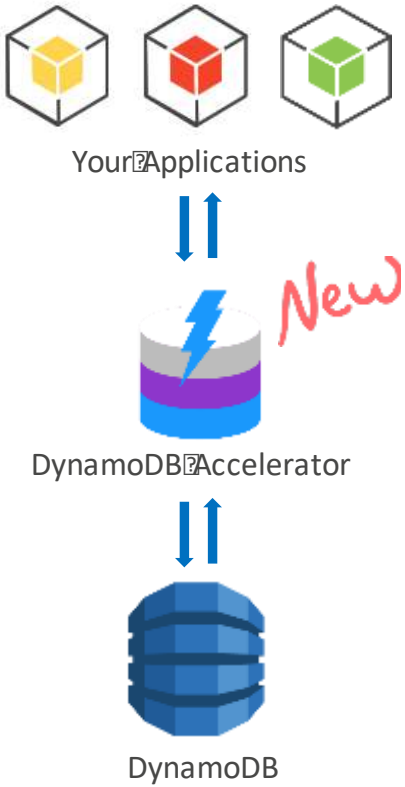
Single-digit millisecond latency

Massive and seamless scalability

Event-driven programming

DynamoDB Accelerator (DAX) Preview

AWS INNOVATE



Features

- **Fully managed, highly available:** handles all software management, fault tolerant, replication across multi-AZs within a region
- **DynamoDB API compatible:** seamlessly caches DynamoDB API calls, no application re-writes required
- **Write-through:** DAX handles caching for writes
- **Flexible**: Configure DAX for one table or many
- **Scalable**: scales-out to any workload with up to 10 read replicas
- **Manageability:** fully integrated AWS service: Amazon CloudWatch, Tagging for DynamoDB, AWS Console
- **Security:** Amazon VPC, AWS IAM, AWS CloudTrail, AWS Organizations

amazon web services | intel

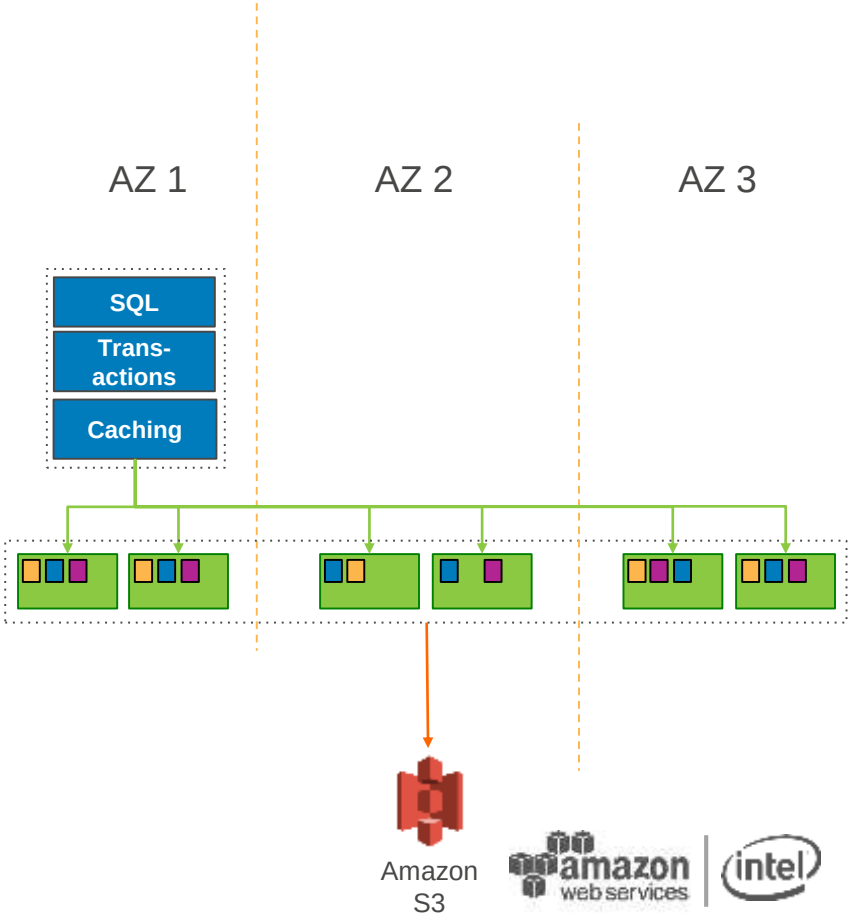
Amazon RDS

- Automated backups (with point-in-time recovery)
- Cross-region snapshot copies
- Automated patch management
- Automated Multi-AZ replication
- Scale up / Scale down instance types
- Scalable storage on demand
- “License included” and BYOL models

Amazon Aurora: MySQL andPostgreSQL-compatible

AWS INNOVATE

- 5x faster than MySQL on same hardware
- SysBench: 100 K writes/sec and 500 K reads/sec
- Designed for 99.99% availability
- 6-way replicated storage across 3 AZs
- Scale to 64 TB and 15 Read Replicas



Amazon Elasticsearch Service

- Distributed search and analytics engine
- Managed service using Elasticsearch and Kibana
- Fully managed – zero admin
- Highly available and reliable
- Tightly integrated with other AWS services

elasticsearch

kibana

Which Data Store Should I Use?

AWS INNOVATE

Data structure

- Fixed schema → SQL, NOSQL
- Schema-Free (JSON) → NoSQL, Search
- (Key, Value) → In-Memory, NoSQL

Access patterns → Store data in the format you will access it

- Put/Get (Key, Value) → In-Memory, NoSQL
- Simple relationships (1:N, M:N) → NoSQL
- Complex relationships (Multi-table joins, transactional) → SQL
- Faceting, Search → Search

Data characteristics → Hot, warm, cold

Cost → Right cost

amazon web services | intel

Analytics Types & Frameworks

PROCESS / ANALYZE

AWS INNOVATE

Batch

- Takes minutes to hours
- Example: Daily/weekly/monthly reports
- Amazon EMR (MapReduce, Hive, Pig, Spark)

Interactive

- Takes seconds
- Example: Self-service dashboards
- Amazon Redshift, Amazon Athena,
- Amazon EMR (Presto, Spark)

Message

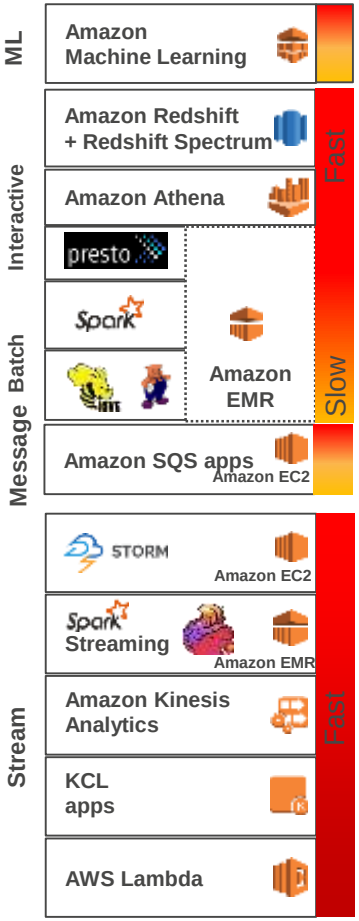
- Takes milliseconds to seconds
- Example: Message processing
- Amazon SQS applications on Amazon EC2

Stream

- Takes milliseconds to seconds
- Example: Fraud alerts, 1 minute metrics
- Amazon EMR (Spark Streaming, Flink), Amazon Kinesis Analytics, KCL, Storm, AWS Lambda

Machine Learning

- Takes milliseconds to minutes
- Example: Fraud detection, forecast demand
- Amazon ML, Amazon EMR (Spark ML)



amazon web services | intel

What About ETL?

AWS INNOVATE

Data Integration Partners

Reduce the effort to move, cleanse, synchronize, manage, and automatize data related processes.

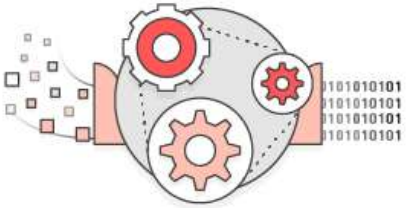
alteryx | ATTUNITY | informatica | ironSource .atom

MATILLION | snapLogic | byte | talend

alooma • • • • •

https://aws.amazon.com/big-data/partner-solutions/

AWS Glue Preview



AWS Glue is a fully managed ETL service that makes it easy to understand your data sources, prepare the data, and move it reliably between data stores

amazon web services | intel

Data Consumption

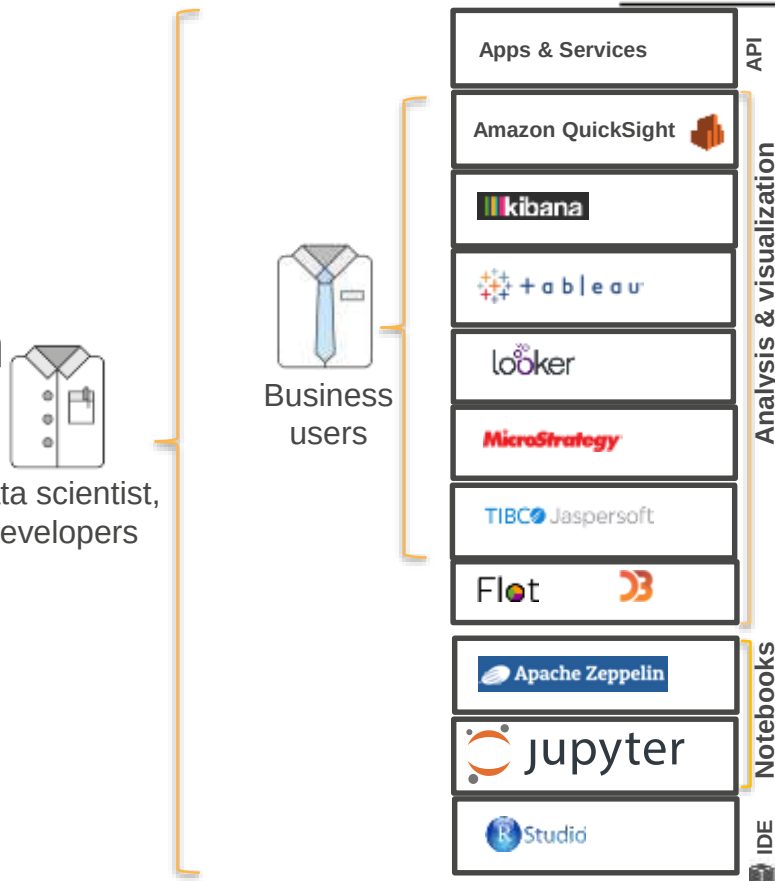
AWS INNOVATE

Applications & API

Analysis and visualization

Notebooks

IDE



amazon web services | intel

Design Patterns and Customer Examples

Primitive: Decoupled Data Bus

AWS INNOVATE

Storage decoupled from processing

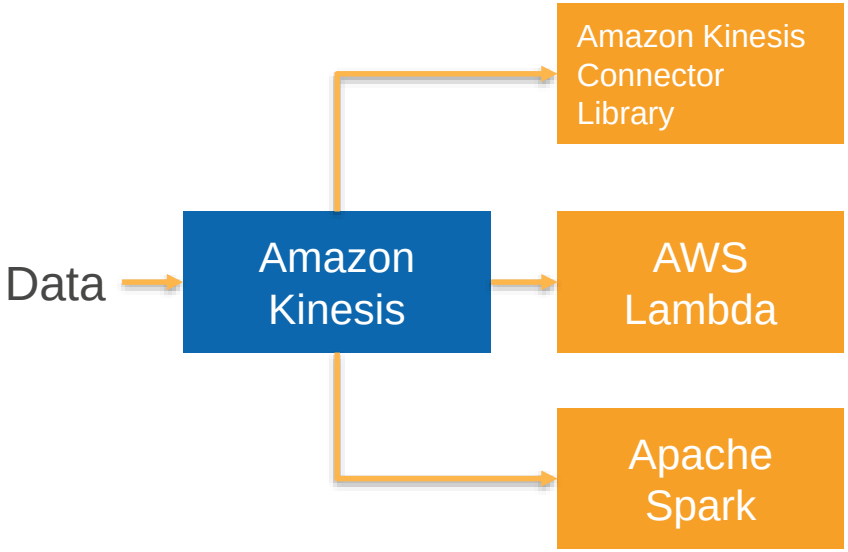
Multiple stages

amazon web services | intel

Primitive: Pub/Sub

AWS INNOVATE

Parallel stream consumption/processing



process

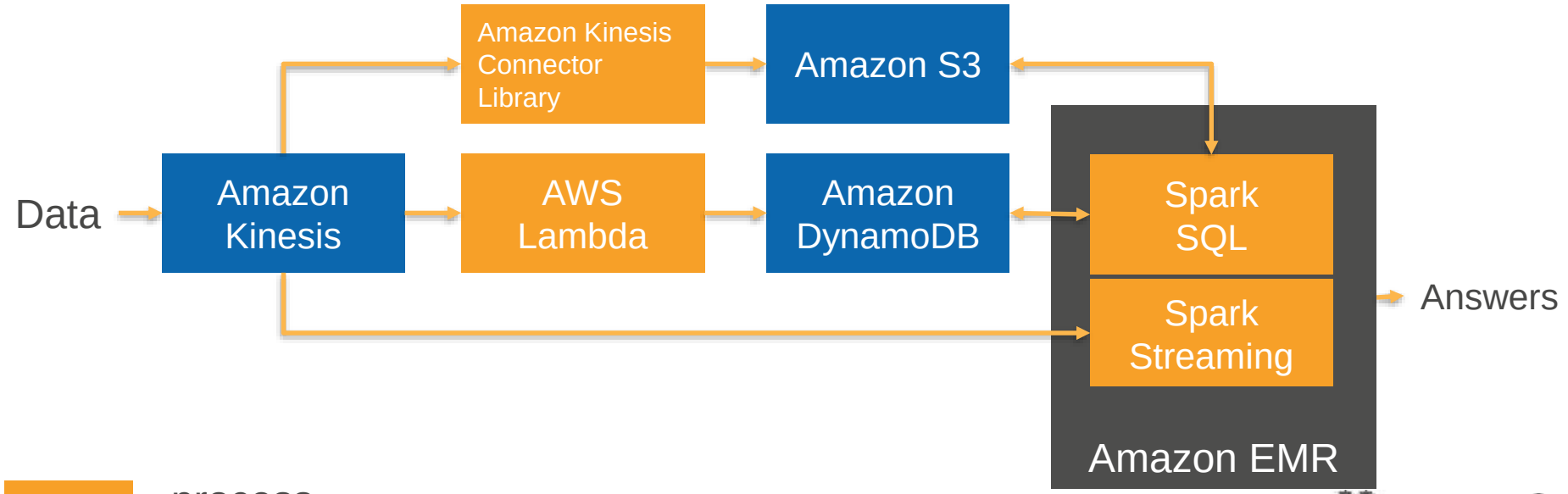
store

amazon web services | intel

Primitive: Materialized Views

AWS INNOVATE

Analysis framework reads from or writes to multiple data stores



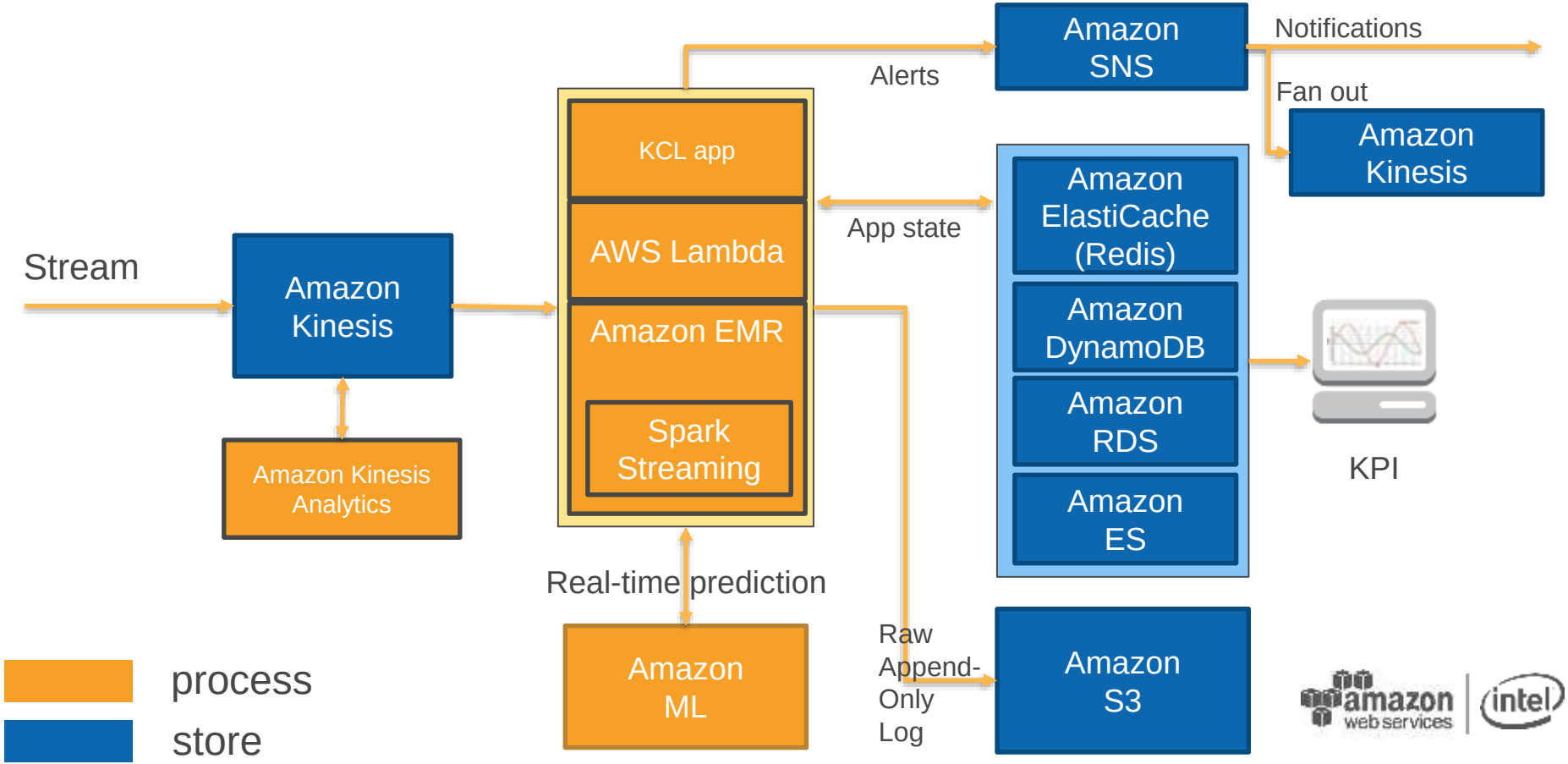
process

store

amazon web services | intel

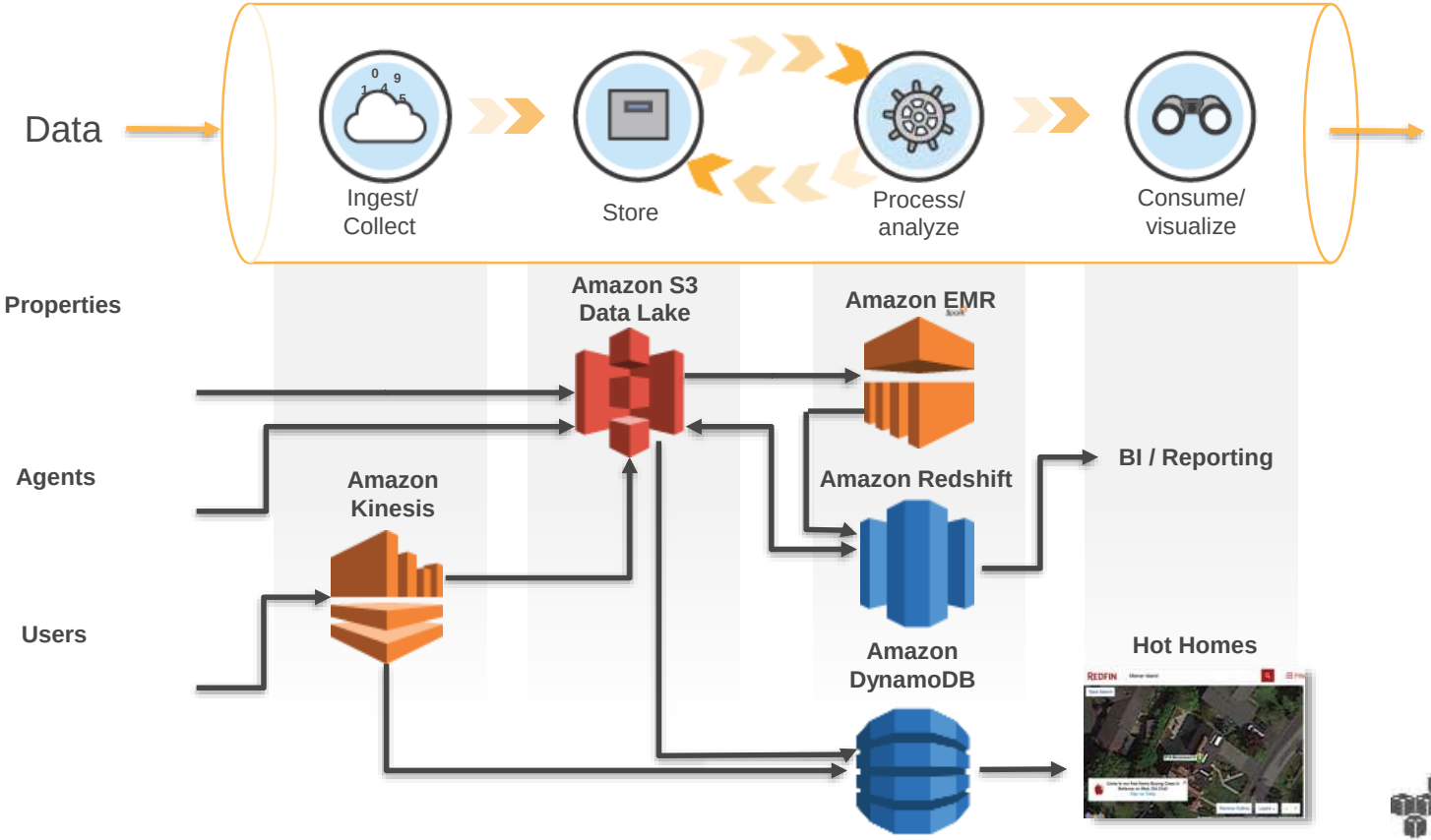
Real-time Analytics

AWS INNOVATE



REDFIN

AWS INNOVATE



User Profile
Recommendation

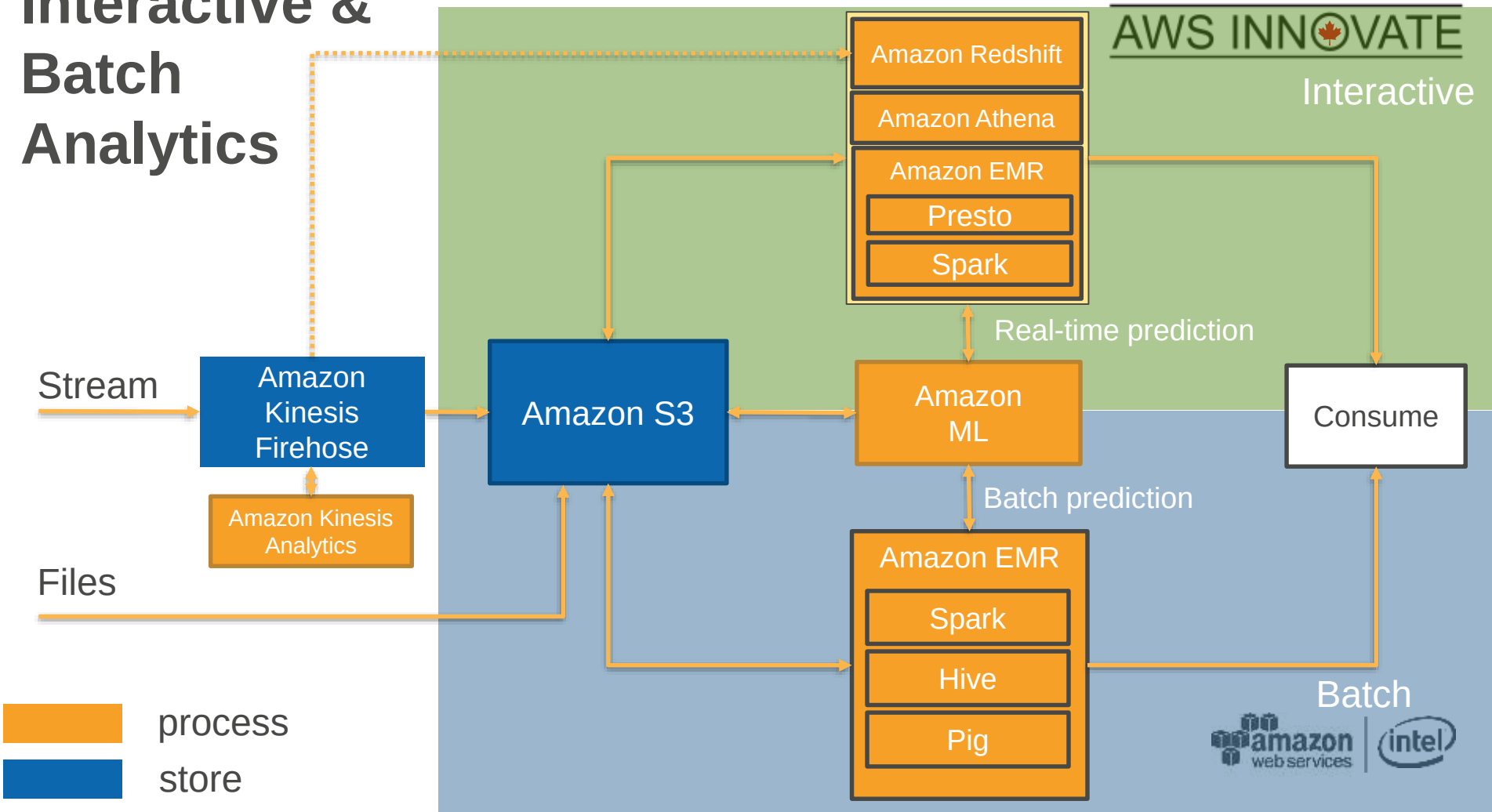
Hot Homes
Similar Homes

Agent Follow-up
Agent Scorecard

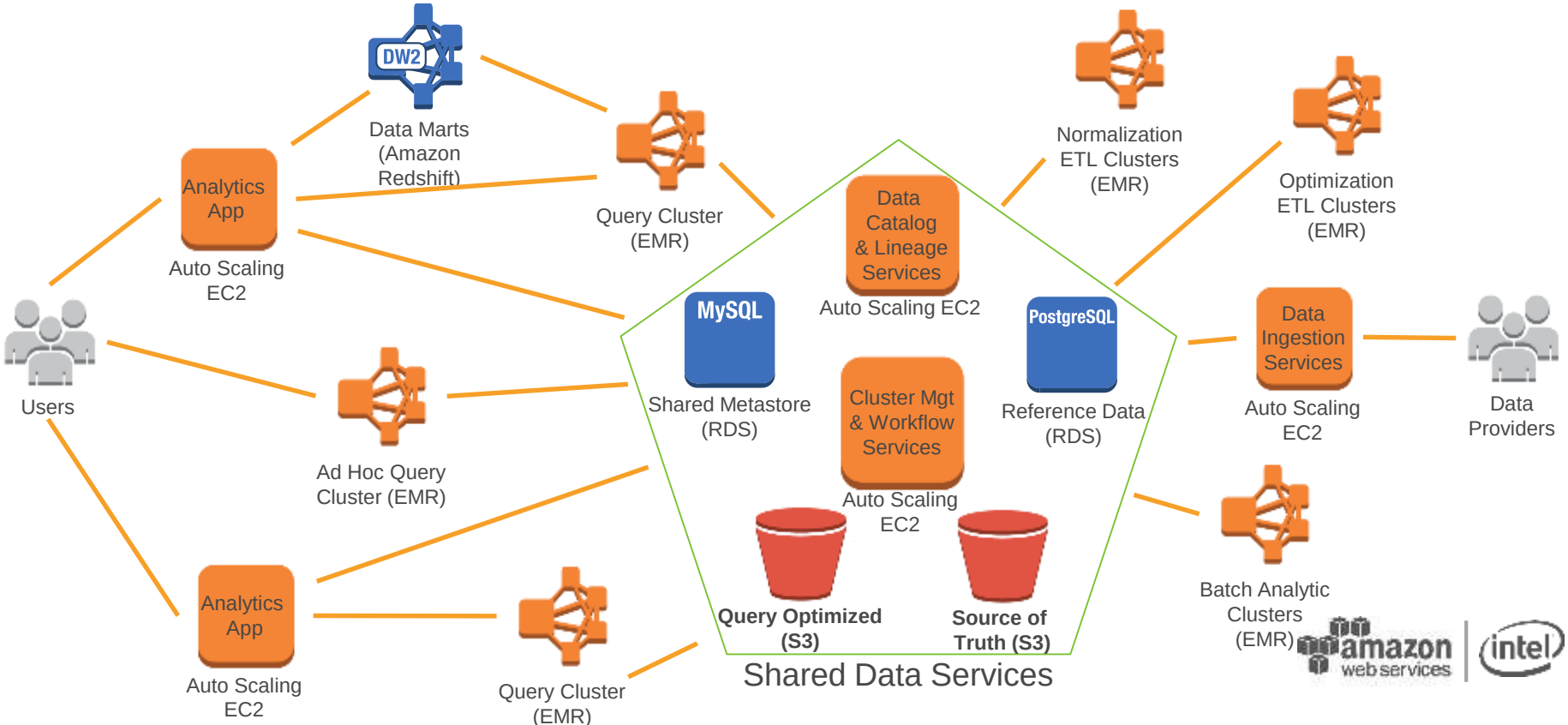
Marketing
A/B Testing
Real Time Data
…

amazon web services | intel

Interactive & Batch Analytics

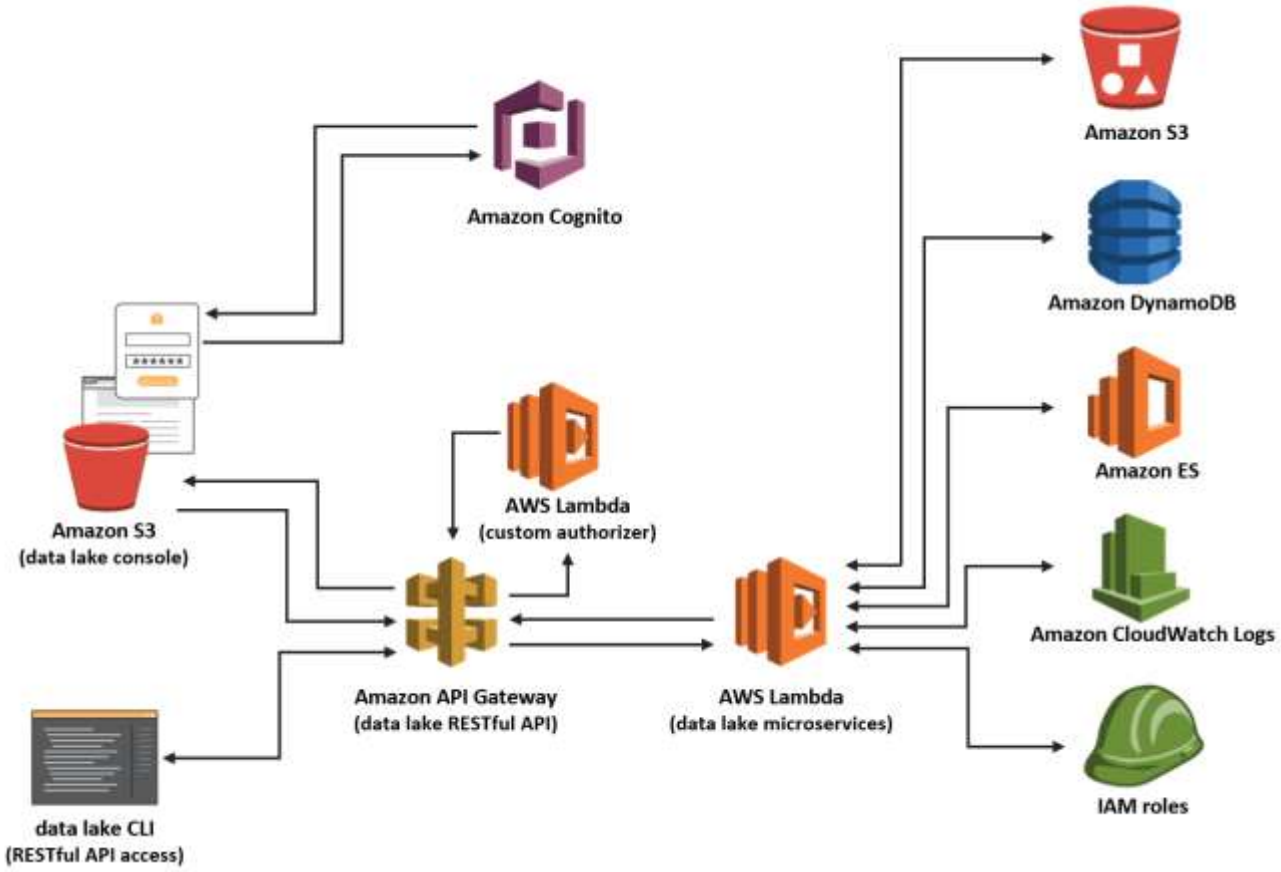


>5 PB, up to 75 billion events per day



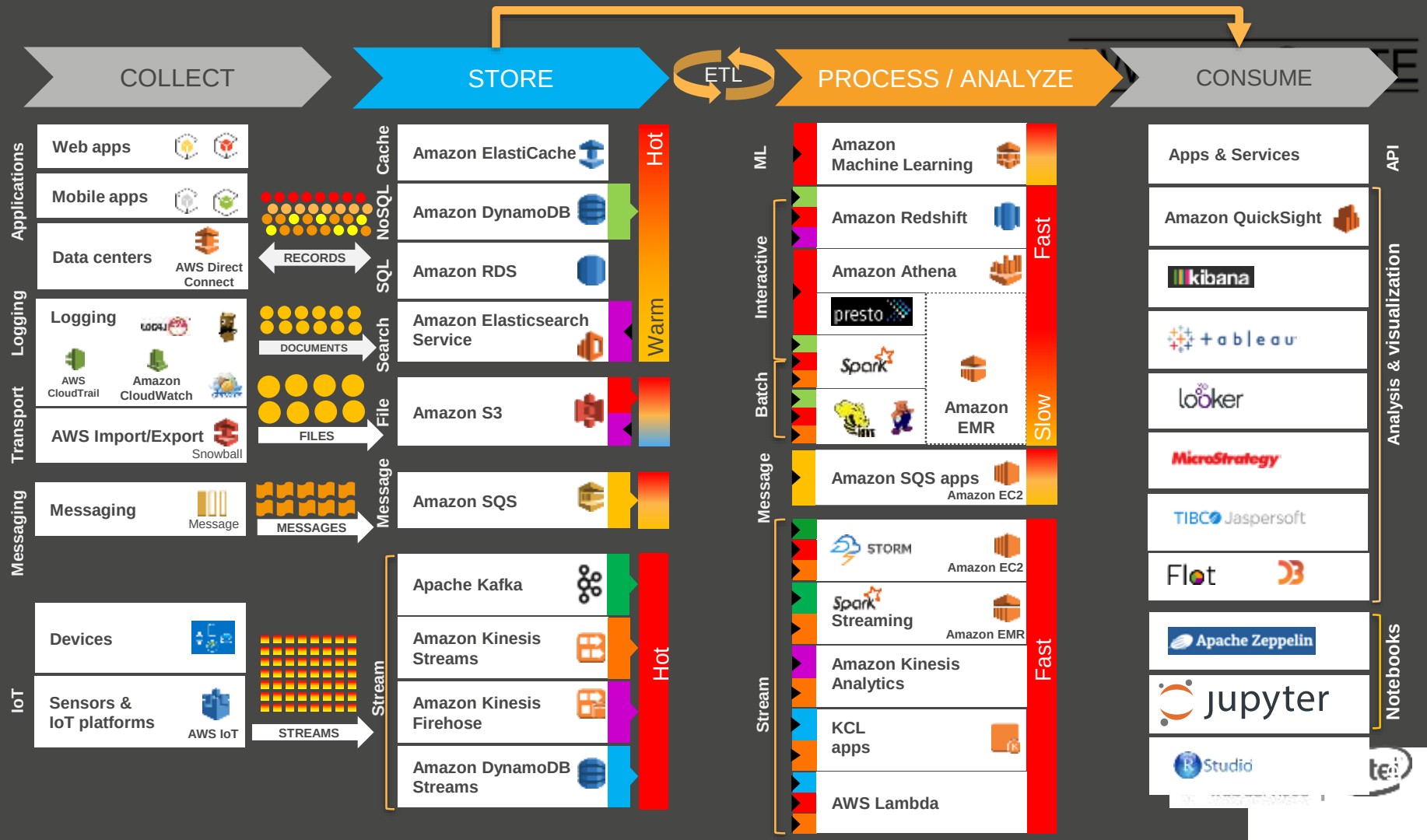
Data Lake

AWS INNOVATE



https://aws.amazon.com/answers/big-data/data-lake-solution/

Putting It All Together


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
IoT
Devices
Sensors & IoT platforms
AWS IoT
RECORDS
DOCUMENTS
FILES
MESSAGES
STREAMS
Cache
Amazon ElastiCache
NoSQL
Amazon DynamoDB
SQL
Amazon RDS
Search
Amazon Elasticsearch Service
File
Amazon S3
Message
Amazon SQS
Stream
Apache Kafka
Amazon Kinesis Streams
Amazon Kinesis Firehose
Amazon DynamoDB Streams
Hot
Warm
Hot
ML
Amazon Machine Learning
Interactive
Amazon Redshift
Amazon Athena
presto
Batch
Spark
Amazon EMR
Message
Amazon SQS apps
Amazon EC2
Stream
STORM
Amazon EC2
Spark Streaming
Amazon EMR
Amazon Kinesis Analytics
KCL apps
AWS Lambda
Fast
Slow
Fast
API
Apps & Services
Analysis & visualization
Amazon QuickSight
kibana
tableau
looker
MicroStrategy
TIBCO Jaspersoft
Flot
Notebooks
Apache Zeppelin
jupyter
Studio

Architectural Principles

AWS INNOVATE

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 (schema-on-read)

Be cost-conscious

- Big data ≠ big cost

amazon web services | intel

Thank you!

Antoine Généreux, AWS Solutions Architect