



# Introduction to Cloud Computing and Amazon Web Services (AWS)

MSBA 6330 Prof Liu

## Outline

- Concepts of cloud computing and virtualization/containerization
- Introduction to (AWS) and key components
- AWS big data line up and use cases
- Comparison between cloud computing platforms

Introduction to Cloud Computing and AWS

## CONCEPTS OF CLOUD COMPUTING, VIRTUALIZATION, AND CONTAINERIZATION

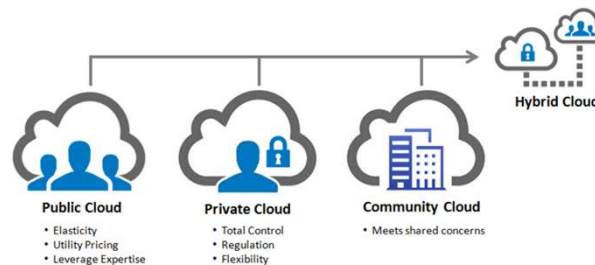
### Cloud Computing

- **Cloud computing** is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.
- Things you can do using cloud computing, e.g.
  - Create new apps and services
  - Store, back up, and recover data
  - Host websites and blogs
  - Stream audio and video
  - Deliver software on demand
  - Analyze data for patterns and make predictions

Source: NIST (<http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf>)

## Types of Cloud Deployment

- **Public:** Provisioned for general public use based on shared physical hardware, owned and operated by a third party provider (Eg. AWS, Azure, Google Cloud).
- **Private:** Provisioned for the use of a single entity, hosted on-site or in a service provider's data center.
- **Hybrid:** A combination of public and private cloud. Use the public cloud for non-sensitive operations, and the private cloud for business-critical operations.



<https://www.rackspace.com/en-us/cloud/cloud-computing/difference>

## Types of Cloud Computing Services

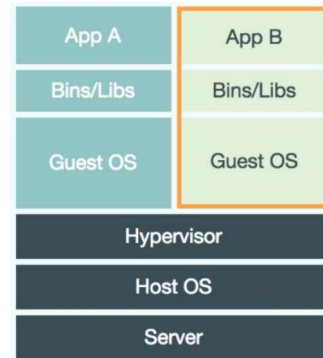
- Cloud computing is commonly characterized as providing three types of functionalities, referred to *IaaS*, *PaaS*, *SaaS*
  - **Infrastructure as a Service (IaaS):** Offer basic building blocks of computing: processing, network connectivity, and storage. Virtual machines, storages, servers, network services.
    - Examples: Amazon's EC2, Microsoft Azure, Rackspace, Google Compute Engine
  - **Platform as a Service (PaaS):** provide a platform for developers to quickly build customized applications. Speed up development.
    - Examples: Google App Engine, Microsoft Azure Web Sites, Force.com
  - **Software as a Service (SaaS):** uses the web to deliver applications that are managed by a third-party vendor (e.g. bug fixes, upgrades, back-end data management) and used by clients.
    - Google Apps, Concur, WebEx, Microsoft Office 365

## Technology behind Cloud computing: Virtualization

- **Virtualization** to make something that doesn't actually (physically) exist appear to exist.

- Microprocess virtualization
- Virtual Memory
- Network virtualization
- **Server Virtualization** (Virtual Machine)

<https://www.youtube.com/watch?v=hPkEqOoQSs4>



Server virtualization

<http://computer.howstuffworks.com/server-virtualization.htm/printable>

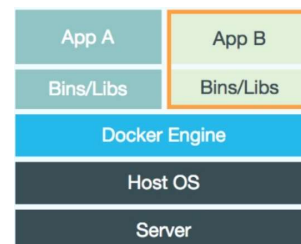
## Containerization - A Step Beyond



- **Container:** A software process whose access has been reduced to the point that it thinks it is the only thing running.
- **Containerization** (or operating-system level virtualization) allows the existence of multiple isolated user-space instances.
  - Docker: a platform for managing containers, started in March 2013.

| Virtualization                       | Containerization                              |
|--------------------------------------|---|
| Virtual Machines (VMs)               | Containers                                    |
| Hardware level virtualization        | Operating system virtualization               |
| Heavyweight                          | Lightweight                                   |
| Slow provisioning                    | Real-time Provisioning and scalability        |
| Limited performance                  | Native performance                            |
| Fully isolated and hence more secure | Process-level isolation and hence less secure |

<https://jaxenter.com/containerization-vs-virtualization-docker-introduction-120562.html>



## Kubernetes – Containers meet clusters

- Kubernetes is a platform hosting Docker containers in a clustered environment with multiple Docker hosts.
  - A container platform
  - A microservices platform
    - A software component of a system that is independently releasable and independent scalable from other parts of the system
  - A portable cloud platform, and a lot more
- Open-sourced by Google in 2014



**Kubernetes**

<https://www.youtube.com/watch?v=R-3dfURb2hA>

<https://www.slideshare.net/imesh/an-introduction-to-kubernetes>

<https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>

Introduction to Cloud Computing and AWS

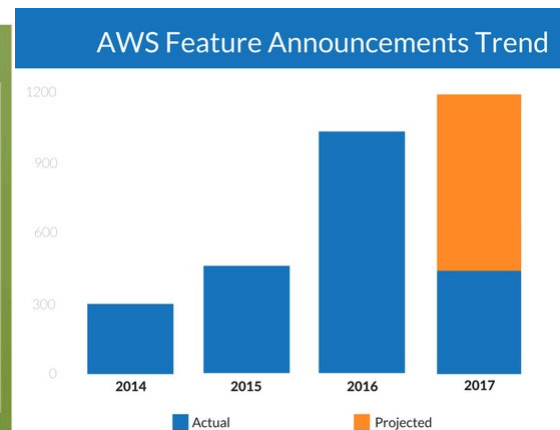
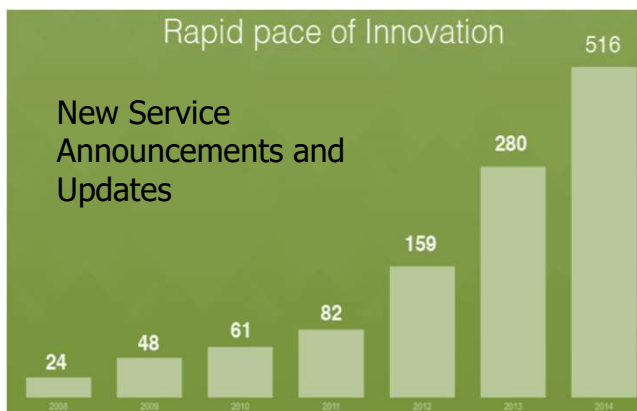
## INTRODUCTION TO AWS AND KEY COMPONENTS

## Amazon Web Services (AWS)

- Grew out of Amazon's need to rapidly provision and configure machines for its own business.
- Early 2000s – Both private and shared data centers began using virtualization to perform “server consolidation”
- 2003 – Internal memo by Chris Pinkham describing an “infrastructure service for the world.”
- 2006 – S3 first deployed in the spring, EC2 in the fall
- 2008 – Competition heds up as Google and Microsoft launch cloud services
- 2009 – Virtual Private Cloud, Relational Database Service,
- 2012 – First customer event Re:invent (30000 registrations). RedShift, DynamoDB
- 2013 – CIA picks AWS over IBM for private cloud
- 2015-16 – Snowball (50 Terabyte appliance) Snowmobile (18-wheel truck with hard drives, 100TB).
- 2016 – AWS surpasses \$10 billion revenue target
- 2017 – AWS nears 100 services (serverless computing, ML etc)

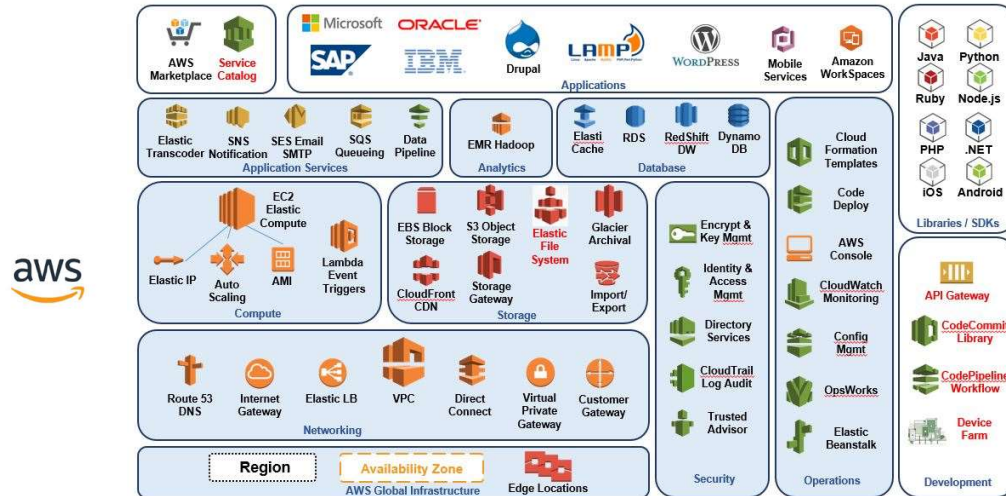
For more details: <https://goo.gl/u3YN5E>

## Rapid Pace of Innovation at AWS



<https://goo.gl/fKXK95>

Today (2015) it looks like this....



For more: <https://goo.gl/1iX2F3>

Source: <https://cloudit4you.wordpress.com/>

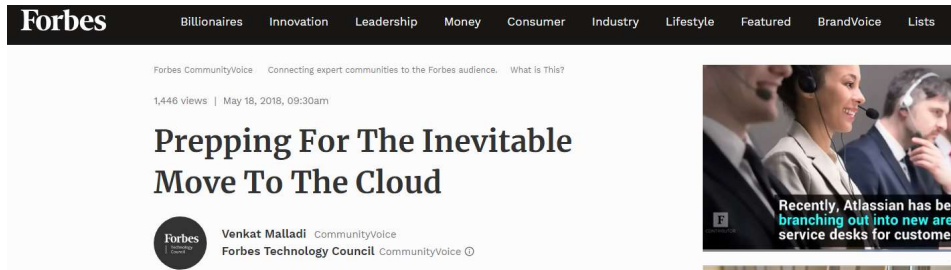
## The tremendous growth of the public cloud industry

- AWS still dominates the cloud market, but Azure, Google and Alibaba are growing faster.

| Cloud vendor                  | Annualized revenue     | % of market | Year-over-year growth |
|-------------------------------|------------------------|-------------|-----------------------|
| Amazon Web Services           | \$18.34 billion        | 51%         | 42%                   |
| Microsoft Azure               | \$6.17 billion         | 17%         | 89%                   |
| IBM Cloud                     | \$4.03 billion         | 11%         | 22%                   |
| Google Cloud Platform         | \$2.05 billion         | 6%          | 125%                  |
| Alibaba Cloud                 | \$1.79 billion         | 5%          | 92%                   |
| Salesforce                    | \$1.78 billion         | 5%          | 31%                   |
| Oracle Cloud                  | \$1.59 billion         | 4%          | 20%                   |
| <b>Subtotal</b>               | <b>\$35.75 billion</b> | <b>86%</b>  | <b>54%</b>            |
| <b>Total Gartner estimate</b> | <b>\$41.79 billion</b> | <b>100%</b> | <b>33%</b>            |

Source (2018): <http://aclouda.com/blog/ms/aws-still-dominating-cloud-market-but-azure-google-and-alibaba-are-growing-faster/>

## The Inevitable Move to the Cloud



- Over 50% of global enterprise in 2018 will rely on public cloud technologies in implementing digital transformation and drive customer experience.

Source: Forbes 2018 <https://goo.gl/hMo1Su>

## AWS Infrastructure

- Characteristics of AWS
  - **Auto Scaling and Elastic load balancing**
    - Scale up or down based on demand
  - Deploy systems in multiple regions
    - Lower latency and better experiences
  - Tools to run a wide range of applications
- **Regions:** physical location in the world
  - Contains multiple availability zones
- **Availability Zones (Azs)**
  - One or more discrete data centers
  - Redundant power/networking/connectivity
  - Housed in separate facilities.

[AWS Cloud Practitioner Essentials](#)

Free training self-paced course



## Compute

- **EC2 = Elastic Compute Cloud**
- Plain English: Amazon Virtual Servers
  - Allow you to obtain and boot new server instances in minutes
  - Allow you to quickly scale capacity, both up and down, as your computing requirements change.
  - Reserved and Spot instances

### Amazon EC2

Elastic Virtual servers  
in the cloud



Amazon EC2

| Model      | vCPU | Mem (GiB) | SSD Storage (GB) |
|------------|------|-----------|------------------|
| m3.medium  | 1    | 3.75      | 1 x 4            |
| m3.large   | 2    | 7.5       | 1 x 32           |
| m3.xlarge  | 4    | 15        | 2 x 40           |
| m3.2xlarge | 8    | 30        | 2 x 80           |

## Storage

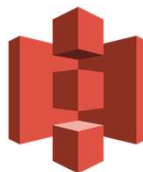
- **S3: Simple Storage Service**
  - In Plain English: Amazon's unlimited FTP server that allows you to store files, images, etc.
  - Scalable, secure, highly available, durable.
  - Used for static content, data for analytics, and back up.

### Amazon S3

Internet scale  
storage via API



Images  
Videos  
Files  
Binaries  
Snapshots



Amazon  
S3

## Simple Storage Service (S3)

- S3 stores objects in **buckets**.
  - A bucket can hold any number of **objects**, which are files of up to 5TB. A bucket has a name that must be **globally unique**.
  - A bucket has a **flat directory structure** (despite the appearance given by the interactive web interface.)

- S3 Addresses

- Internal address (used within AWS): e.g.
  - `s3://mybucket1245/auction_data/auctiondata.csv`
- External address (if made public)
  - [https://s3.amazonaws.com/mybucket1245/auction\\_data/auctiondata.csv](https://s3.amazonaws.com/mybucket1245/auction_data/auctiondata.csv)

Bucket  
name

Object  
name



## Storage

- **EBS - Elastic Block Store**
- In plain English – virtual hard disk to be used by EC2 instances.
  - Low-latency performance & scalability
  - More expensive than S3 storage
- Also: **Amazon Glacier** (for low-cost archive).
  - Optimized for infrequent access.
  - Extremely low cost (\$0.01/GB/Month).

**Amazon EBS**  
Block storage for use  
with Amazon EC2

EC2



**Amazon Glacier**  
Storage for archiving  
and backup



Images  
Videos  
Files  
Binaries  
Snapshots

## Database

### • RDS – Relational Database Service

- Fully managed database services
- MySQL, Oracle, SQL Server, PostgreSQL, & Amazon Aurora (Amazon's MySQL Replacement).
- Autoscaling
- Support backup



Amazon  
RDS

### Amazon RDS

Managed relational  
database service



## Database

### • DynamoDB – NoSQL keystore database

- Like Amazon's MongoDB
- Milliseconds latency

| User ID<br>(Hash Key) | User Segment<br>(Range Key) | Timestamp<br>(Attribute) |
|-----------------------|-----------------------------|--------------------------|
| 1234                  | Segment1                    | 1448895406               |
| 1234                  | Segment2                    | 1448895322               |
| 1235                  | Segment1                    | 1448895201               |



Amazon  
DynamoDB

### • ElastiCache - In-memory caches for fast performance

- Support engines: Memcached & Redis



Amazon  
ElastiCache

## Management Tools

- **IAM** – Identity and Access Management – users, permission, SSH keys etc.



- **VPC: virtual private cloud**

- In Plain English: Make all of your AWS services are on the same little network (separated from other things on AWS). Amazon's VLANs.



Amazon VPC

### AWS IAM (Identity & Access Mgmt)

Manage users, groups & permissions



### Amazon VPC:

Private, isolated section of the AWS Cloud



## More Information on AWS Services and Categories

- Amazon Virtual Private Cloud (VPC)
- Security Groups
- EC2
- Lambda
- AWS Elastic Beanstalk
- Auto Scaling
- Amazon Elastic Block Store (EBS)
- Amazon Simple Storage Service (S3)
- Amazon Glacier
- Amazon RDS
- Amazon DynamoDB

- Amazon RedShift
- Amazon Aurora
- Etc.

AWS Cloud Practitioner Essentials (Intro and demos)

<https://www.aws.training/learningobject/curriculum?id=16357>

Hands on labs:

<https://amazon.qwiklabs.com/catalog?locale=en>

Introduction to Cloud Computing and AWS

## AWS BIG DATA LINE UP AND USE CASES

### AWS's Big Data Line up

- **Kinesis Streams**: store and analyze real-time streaming data (by default, data is retained for 24 hours)
- **AWS Lambda**: serverless computing; let you run code without provisioning or managing servers. Useful real time, event driven processing.
- **EMR (elastic mapreduce)**: amazon's Hadoop stack, including Hive, Pig, spark, etc. Ideal for batch ETL, ad hoc analytics and data mining.
- **Amazon Machine Learning**: let you build ML models through wizards and run against data stored in S3, Redshift or RDS. (up to 100GB; bigger datasets should use EMR/Spark MLlib).
- **Amazon DynamoDB**: NoSQL database for low-latency frequent read/write. Commonly used for supporting live apps (mobile apps, voting, ad serving, e-commerce websites, streaming data storage, etc).
- **Amazon RedShift**: petabyte scale data warehouse; SQL-based; Useful for large scale OLAP and BI reporting.
- **Amazon Elasticsearch**: real time distributed full-text search, structured search and analytics.
- **Amazon QuickSight**: (2015) a business intelligence, visualization, ad hoc analysis tool.
- **Amazon EC2**: for building your self-managed big data analytics applications.

[Ref: Big Data Analytics Options on AWS](#)

## Big Data

- **EMR** – Elastic Map Reduce

- Amazon's Hadoop Cluster, auto scaling, fully managed
- Use EC2 as its computing instances.
- Transient versus Persistent modes

Release | emr-5.3.0

|  |  |   |
|--|--|---|
| <input checked="" type="checkbox"/> Hadoop 2.7.3 | <input type="checkbox"/> Zeppelin 0.6.2        | <input type="checkbox"/> Tez 0.8.4      |
| <input type="checkbox"/> Flink 1.1.4             | <input type="checkbox"/> Ganglia 3.7.2         | <input type="checkbox"/> HBase 1.2.3    |
| <input checked="" type="checkbox"/> Pig 0.16.0   | <input checked="" type="checkbox"/> Hive 2.1.1 | <input type="checkbox"/> Presto 0.157.1 |
| <input type="checkbox"/> ZooKeeper 3.4.9         | <input type="checkbox"/> Sqoop 1.4.6           | <input type="checkbox"/> Mahout 0.12.2  |
| <input checked="" type="checkbox"/> Hue 3.11.0   | <input type="checkbox"/> Phoenix 4.7.0         | <input type="checkbox"/> Oozie 4.3.0    |
| <input type="checkbox"/> Spark 2.1.0             | <input type="checkbox"/> HCatalog 2.1.1        |   |

- **Redshift** - Amazon's petabyte-scale data warehouse,
  - Support standard SQL and BI tools
  - Uses columnar storage format.
  - Massively parallel query execution
  - Include Redshift Spectrum for query unstructured data in S3 (including Avro, CSV, ORC, RegexSerDe, textfile etc)

### Amazon EMR (Elastic Map Reduce)

Hosted Hadoop  
framework



### Amazon Redshift

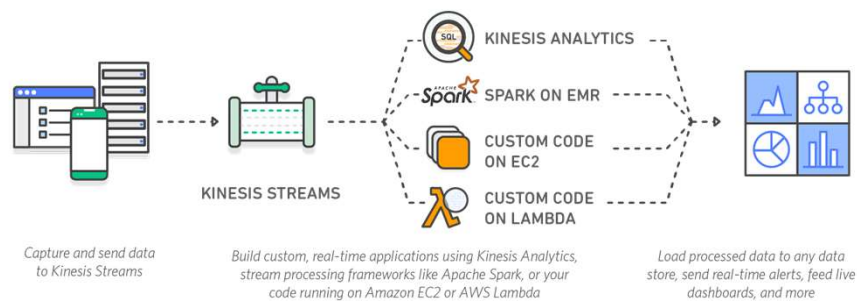
Petabyte-scale data  
warehouse service



## Streaming Processing

- **Kinesis** – for real-time ingestion & processing of streaming data

- Amazon's Kafka, fully managed, scalable
- Ingest, process, and analyze real-time data such as application logs, website clickstreams, IoT telemetry data



See more reference architectures @ [https://aws.amazon.com/kinesis/?nc2=h\\_m1](https://aws.amazon.com/kinesis/?nc2=h_m1)

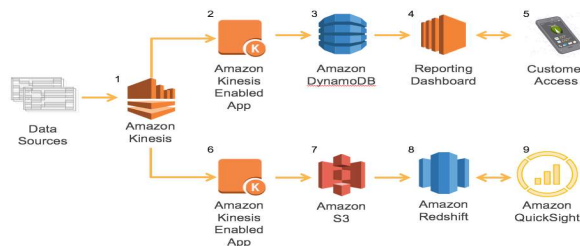
## AWS big data use case 1- Enterprise data warehouse

- Enterprise data warehouse
  - Ingest data into S3
  - Use EMR to transform and cleanse data, store results to S3
  - Redshift loads, sorts, distributes, and compress data into tables for fast OLAP analytics queries.
  - QuickSight can be used for analytics, or external visualization tools connected to RedShift



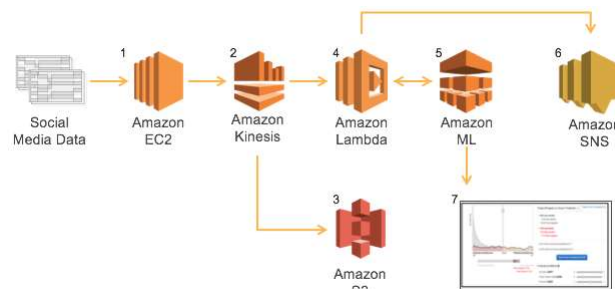
## AWS big data use case 2- Capturing and Analyzing Sensor Data

- Sensor data uploaded into Kinesis streams.
- Custom Kinesis app on EC2 reads data, and send sensitive data to near real-time dashboard reporting (2-5), DynamoDB is used for fast storage. Reporting dashboard is a custom-build web-application using EC2 and data in DynamoDB
- Steps 6-9 process the same data at slower pace for non-real-time data warehouse style analytics and reporting.



### AWS big data use case 3- Sentiment Analysis of Social Media Data

- EC2 instances harvests social media data through APIs. Multiple social media streams are publishes to Kinesis.
- Raw data is stored in S3 for long term archival.
- Lambda is used for processing/normalizing data and request predictions from Amazon ML in near real time. ML also produces performance metrics via AWS console.
- Actionable data (alerts, predictions, via emails or text) is sent to Amazon SNS (simple notification service).



### Popularity of Big Data in the Cloud

- More and more big data deployments are in the public cloud

Apache Spark deployments in the public cloud increased in 2016. In contrast, the percentage of Spark deployments on-premises decreased in the past year.

APACHE SPARK DEPLOYMENT  
IN PUBLIC CLOUDS  
HAS INCREASED BY 10%  
SINCE 2015.



<https://databricks.com/blog/2016/09/27/spark-survey-2016-released.html>



## Use Case – NY Times

### NYTimes “TimesMachine” (June 2008)



1851-1922 Articles

TIFF -> PDF

Input: 11 Million Articles  
(4TB of data)

#### What did he do ?

Spun 100 EC2 Instances for 24 hours

Input: All data on S3

Output: 1.5 TB of Data

Used: Hadoop, iText, JetS3t


## Use Case - Netflix

- Moved its entire technology infrastructure in Nov 2012
- Linux-based Web Servers
  - EC2 instances
  - Use Amazon S3, Cassandra Database services
- Transcoding
  - EMR for computing, S3 for storage
- Recommendation
  - Hive & Pig & Machine Learning for offline analytics and model building

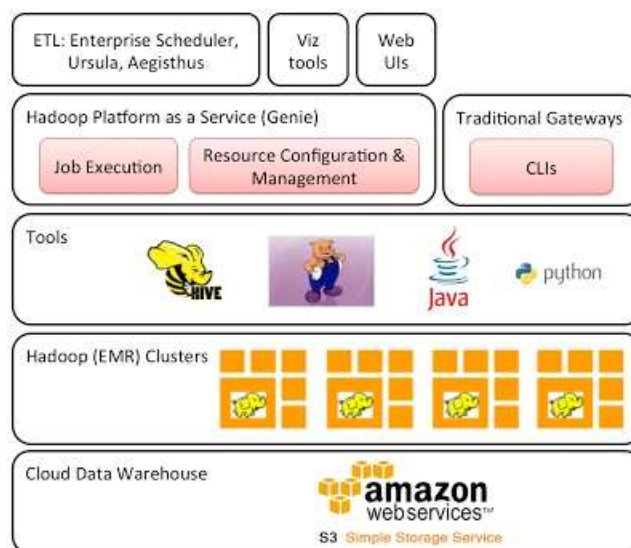
<https://aws.amazon.com/solutions/case-studies/netflix/>

## Why Netflix Uses Cloud Infrastructure

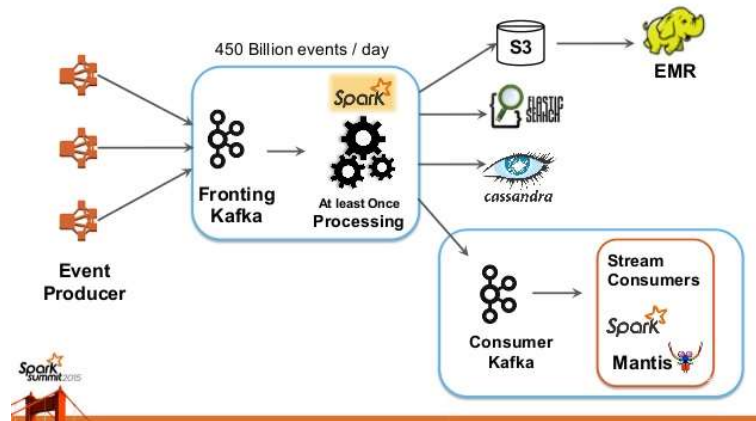
Get stuck with wrong config  
 Wait Wait File tickets  
 Ask permission Wait Wait  
 Wait Things we don't do Wait  
 Run out of space/power  
 Plan capacity in advance  
 Have meetings with IT Wait



## Use Case - Netflix

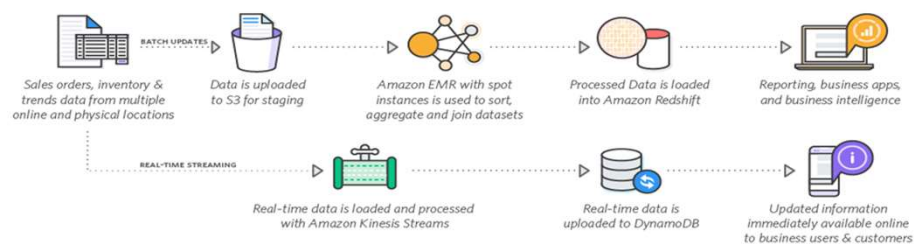


## Use Case – Netflix' Event Data Pipeline



<http://www.slideshare.net/SparkSummit/spark-and-spark-streaming-at-netflix-sedakar-daxini>

## Big Data Cloud Computing Use Case: Lyft



- Ride location tracking
  - DynamoDB (NoSQL)
- Lyft App events
  - open an app, driver action car movement, user action
  - Amazon Kinesis



1 Million service requests/sec

10,000 people in Lyfts at any moment in US

Introduction to Cloud Computing and AWS

## COMPARISON BETWEEN CLOUD COMPUTING PLATFORMS

### Comparisons between Major Cloud Providers: Compute

| Function                            | <a href="#">AWS</a>  | <a href="#">Azure</a>                      | <a href="#">Google</a>                                    |
|-------------------------------------|--|--|---|
| <b>Virtual Machines, Containers</b> | EC2 + EC2 Container Service                                    | Virtual Machines, Azure Kubernetes Service | Compute Engine<br>Kubernetes Engine                       |
| <b>Event driven compute</b>         | Amazon Lambda  | Functions                                  | Cloud Functions   |
| <b>Realtime data processing</b>     | Amazon Kinesis, Amazon data pipeline                           | Event hubs<br>Apache Storm for HDInsight   | Cloud Dataflow  |
| <b>Hadoop</b>                       | Elastic MapReduce  | HDInsight                                  | Cloud Dataproc  |
| <b>Machine Learning</b>             | Amazon Machine Learning, MXNet, Deeep Learning, TensorFlow etc | Azure Machine Learning etc.                | Cloud Machine Learning Engine, Cloud Deep Learning Image, |

<https://www.whizlabs.com/blog/aws-vs-azure-vs-google/>

## Comparisons Between Major Cloud Providers: Compute

| Function                         | AWS                 | Azure                         | Google                    |
|----------------------------------|---------------------|-------------------------------|---------------------------|
| <b>Data storage web services</b> | Amazon S3           | Storage                       | Cloud Storage             |
| <b>Archiving</b>                 | AWS Glacier         | Azure Backup                  | Cloud Storage Nearline    |
| <b>Database as Service</b>       | Amazon RDS , Aurora | Azure SQL Database, Cosmos DB | Cloud SQL                 |
| <b>NoSQL</b>                     | DynamoDB            | Table Storage                 | Cloud Datastore, BigTable |
| <b>Data Warehousing</b>          | AWS Redshift        | SQL Data Warehouse            | BigTable / BigQuery       |

## Demo: Google BigQuery

- Google BigQuery
  - SQL based large-scale cloud data warehouse
  - Google's Answer to Amazon's Redshift
  - Pay per query and for table storage

|                   |                          |
|-------------------|--------------------------|
| Storage           | \$0.02 per GB, per month |
| Long Term Storage | \$0.01 per GB, per month |
| Streaming Inserts | \$0.05 per GB            |
| Queries           | \$5 per TB               |



Reference: <https://cloud.google.com/files/BigQueryTechnicalWP.pdf>

## Resources

- Video tutorials (great resource to get introduced to AWS)
  - <http://aws.amazon.com/getting-started/>
  - [AWS Educate program](#). Free videos and AWS tutorials
- [AWS documentation](#)
  - most comprehensive source of information
- [AWS EMR developer guide](#):
  - handbook of EMR services and how to use them.
- [AWS EMR management guide](#):
  - How to setup, manage, and debug EMR clusters.
- [Big Data Analytics Options on AWS](#):
  - Ideal usage of big data tools on AWS, including redshift, Kinesis, EMR, DynamoDB, ML, Lambda, ElasticSearch and three stylized use cases.
- AWS Tutorials & Training for Big Data
  - <https://aws.amazon.com/big-data/getting-started/tutorials/>
- Data Lakes and Analytics on AWS
  - <https://aws.amazon.com/big-data/datalakes-and-analytics/>

## Resources

- Videos
  - [Google Cloud Platform & Big Data](#)
  - [Google Cloud Platform 2016 Keynote](#)
  - [AWS 2016 Keynote](#)
  - [GOTO Conferences: The latest in software development](#)
- AWS, Google Cloud and Azure links
  - <http://blogs.aws.amazon.com/bigdata/>
  - <https://cloudplatform.googleblog.com/>
  - <https://azure.microsoft.com/en-us/>
  - <https://cloud.google.com/free-trial/>