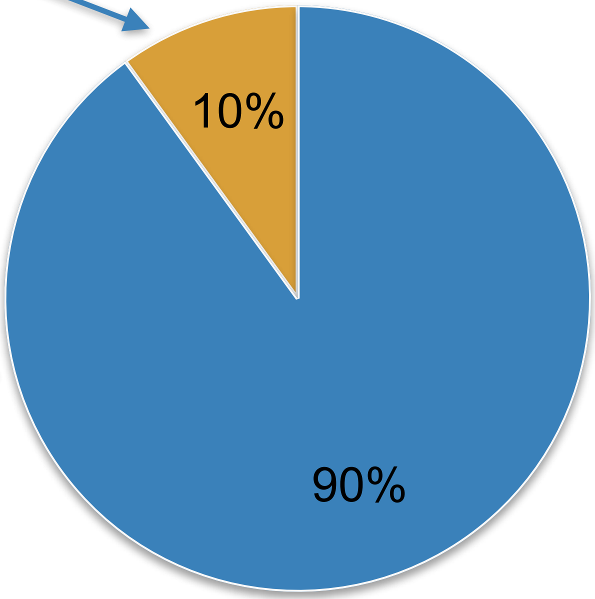
CORNELL
NYCTECH

Data Pipelines

cs5356
Daniel Doubrovkine
@dblockdotorg

Job: Data Scientist



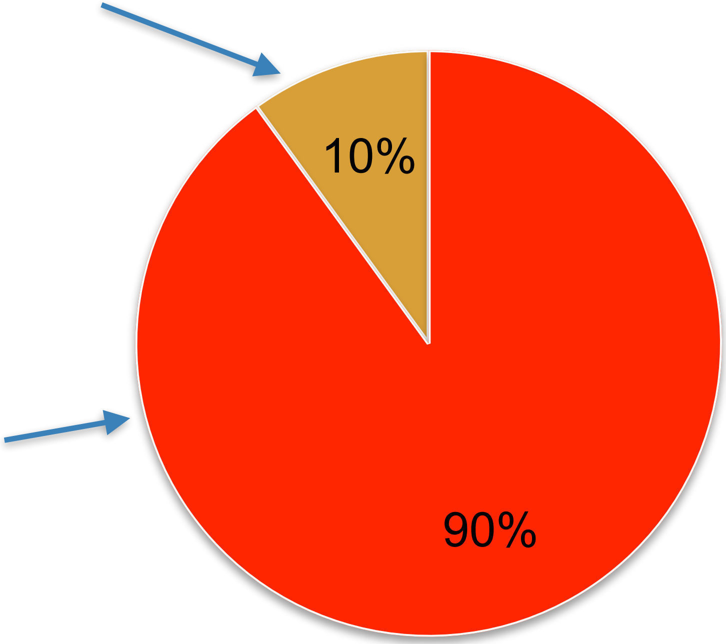
Machine Learning Big Data in the Cloud …

Moving Data Around

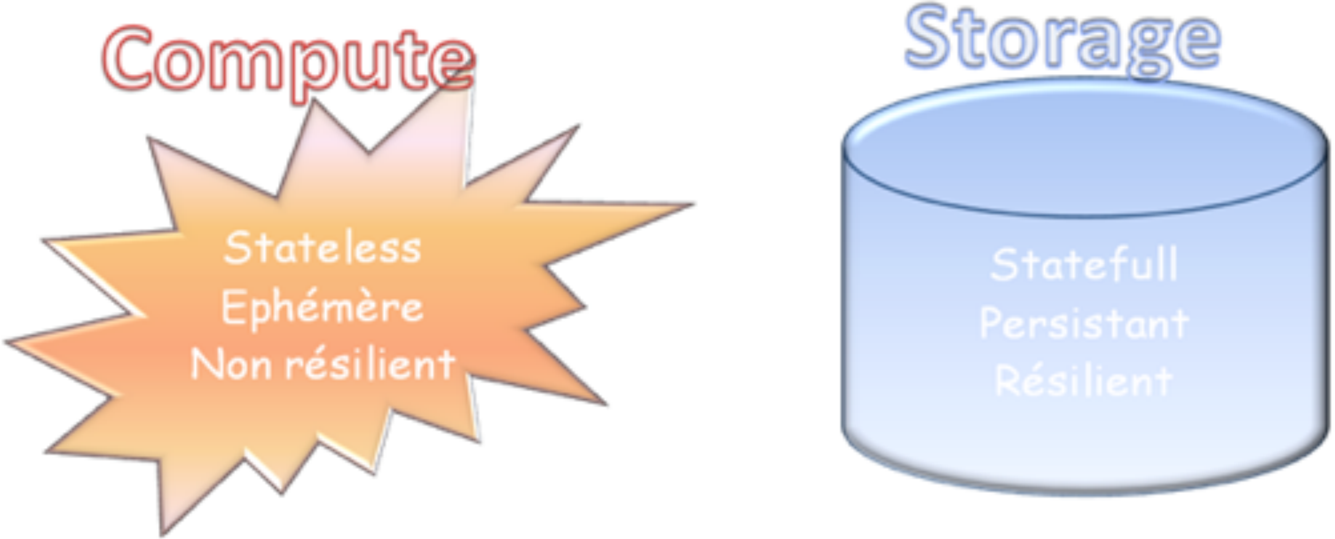
Data Pipelines

Machine Learning Big Data in the Cloud …

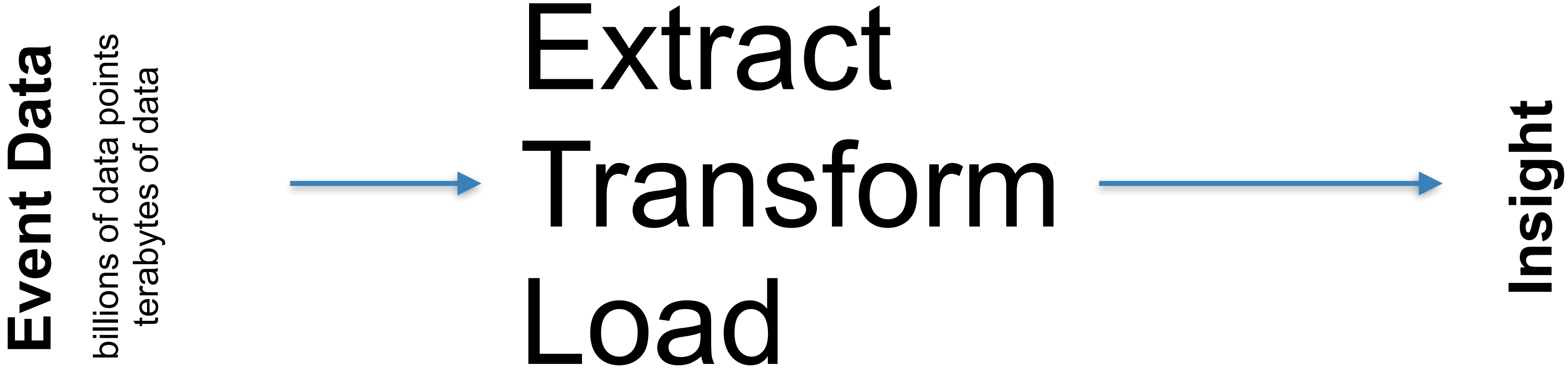
Moving Data Around



Moving Data Around



Typical Warehousing: Batch



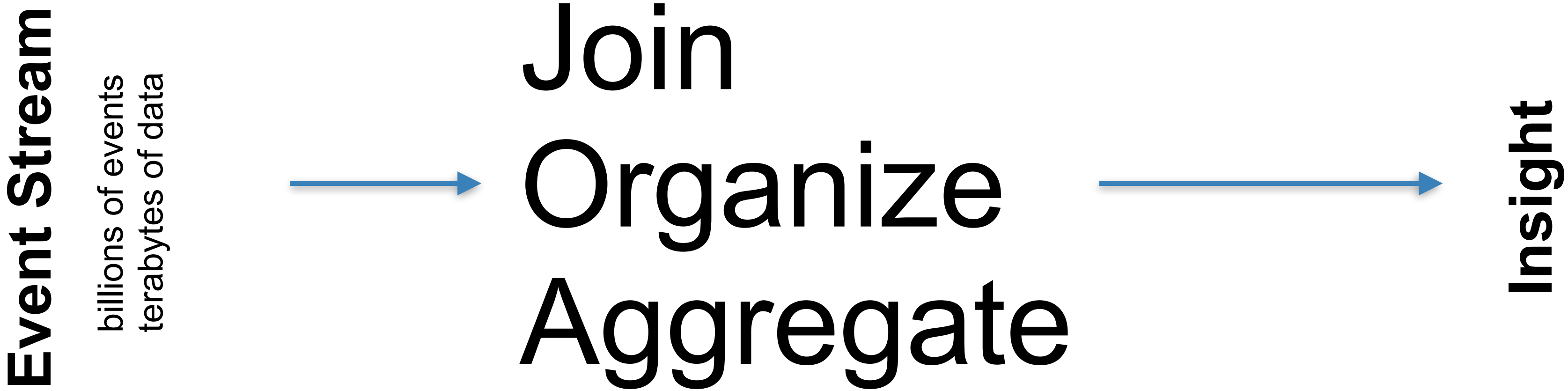
Event Data

billions of data points
terabytes of data

Extract
Transform
Load

Insight

Too slow … continuous processing



Event Stream

billions of events
terabytes of data

Join
Organize
Aggregate

Insight

Clickstream

- What users do, usually in log files.

  date, time, IP, URL

- Page views per URL over time?
- Top N page views of all time?
- What products to visitors buy together?

Aggregate

- Distributed Messaging System
  - **In:** Logs
  - **Out:** Domain Data

kafka

Hadoop

- Scalable
- Reliable
- Fault Tolerant
- Simple

One Big Computer

https://hadoop.apache.org

Hadoop

- Storage: HDFS

  Linux ext3
  Replicated blocks
  Write Once / Read Many
  Retry transfers

- Execution: Map/Reduce

  In parallel on many servers
  Retry on failure
  Map + Reduce
  Servers can join or leave

Map/Reduce Canonical Example

- Word Count
  - **Map**: transform text to { word: 1 }
  - **Reduce**: sum by key

Storage

amazon DynamoDB

amazon webservices™ S3

YMMV

Amazon RDS

APACHE HBASE

Amazon Redshift

mongoDB

Replacing Parts of Hadoop

- Spark
- Microsoft Dryad
- Apache Tez
- Impala
- Google Big Query
- Google Cloud Dataflow

Spark

- Richer API

  filter(), join(), distinct(), groupByKey()

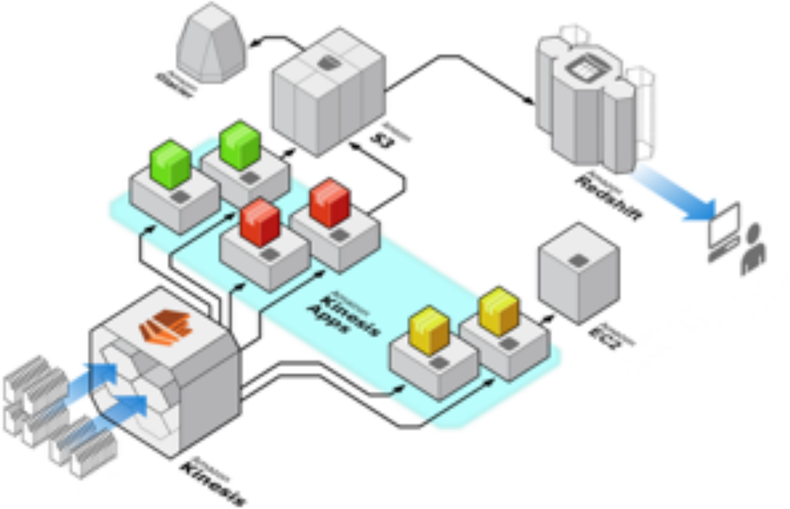
- Maintain State

  updateStateByKey()

- Window Operations

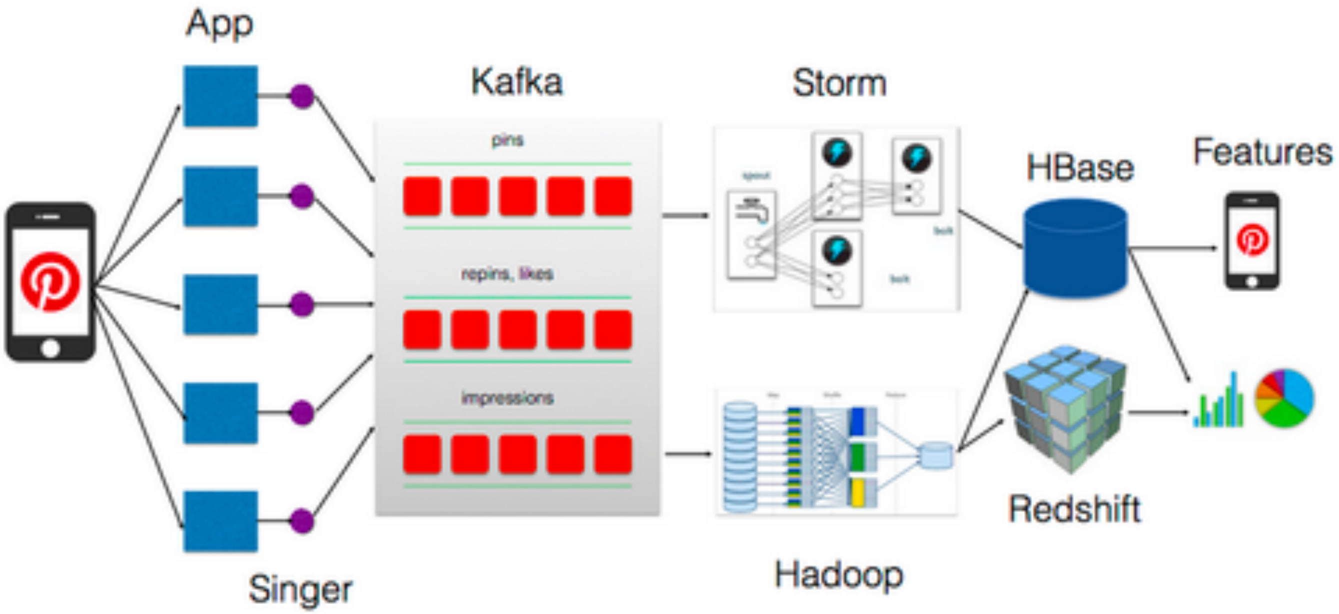
  window size and slide interval

Amazon Kinesis



- Easy
- Real-Time
- High Throughput
- S3/DynamoDB/Redshift

Pinterest

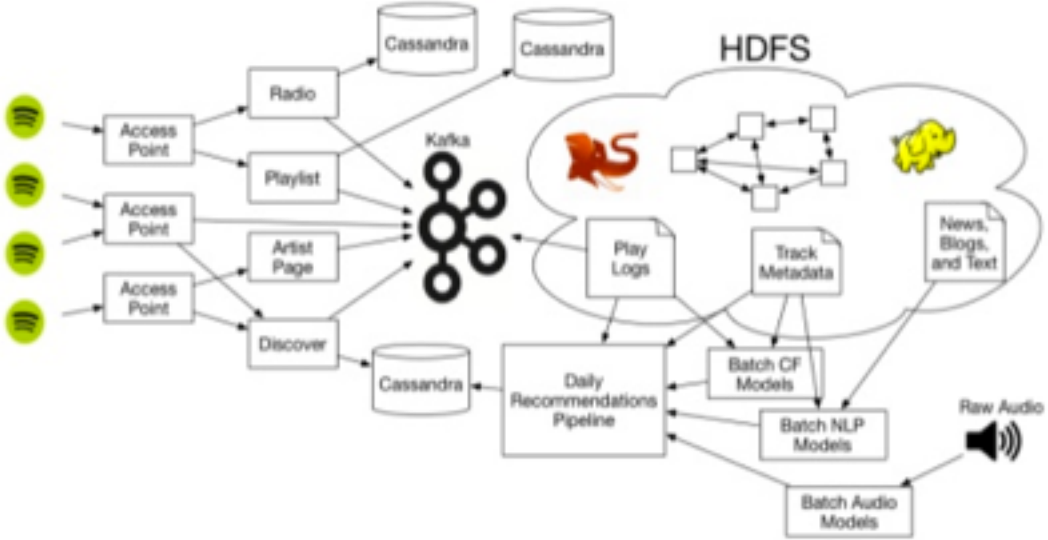


Data Architecture overview

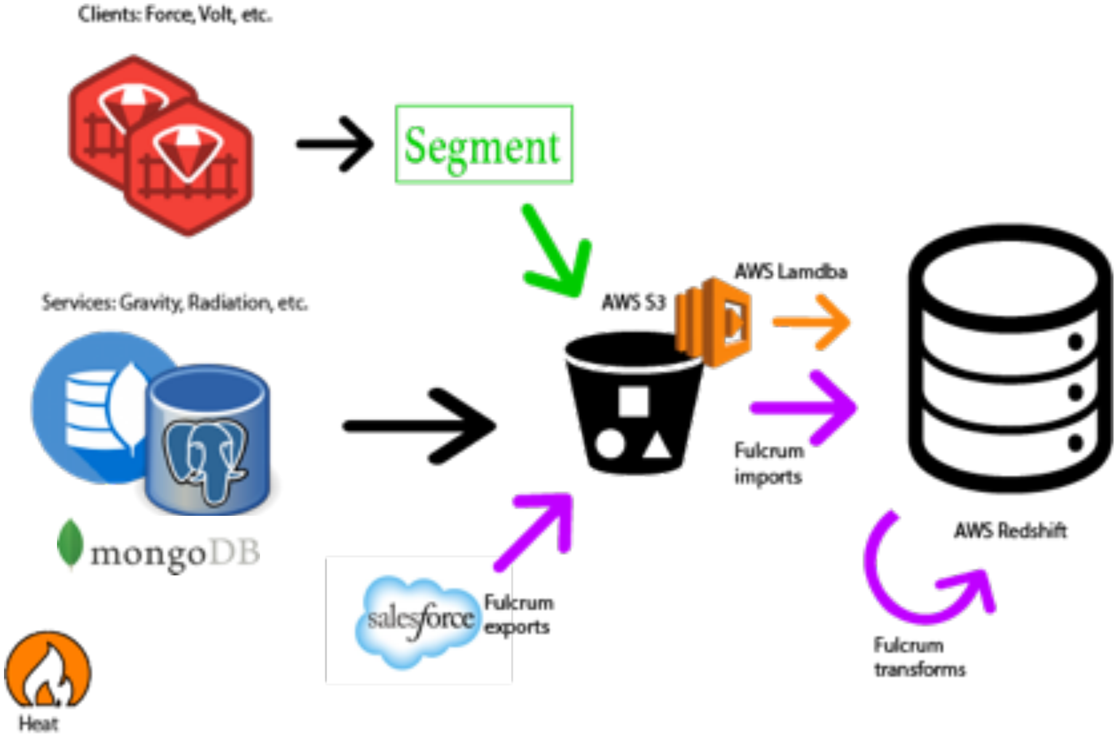
CORNELL NYC TECH

Spotify

Recommendation systems



Artsy


Clients: Force, Volt, etc.
Segment
Services: Gravity, Radiation, etc.
mongoDB
Heat
salesforce
Fulcrum exports
AWS S3
AWS Lamdba
Fulcrum imports
AWS Redshift
Fulcrum transforms

MongoDB (after the break)

Shannon Bradshaw

mongoDB