Data Engineering

April 2019 - STL Big Data User Group

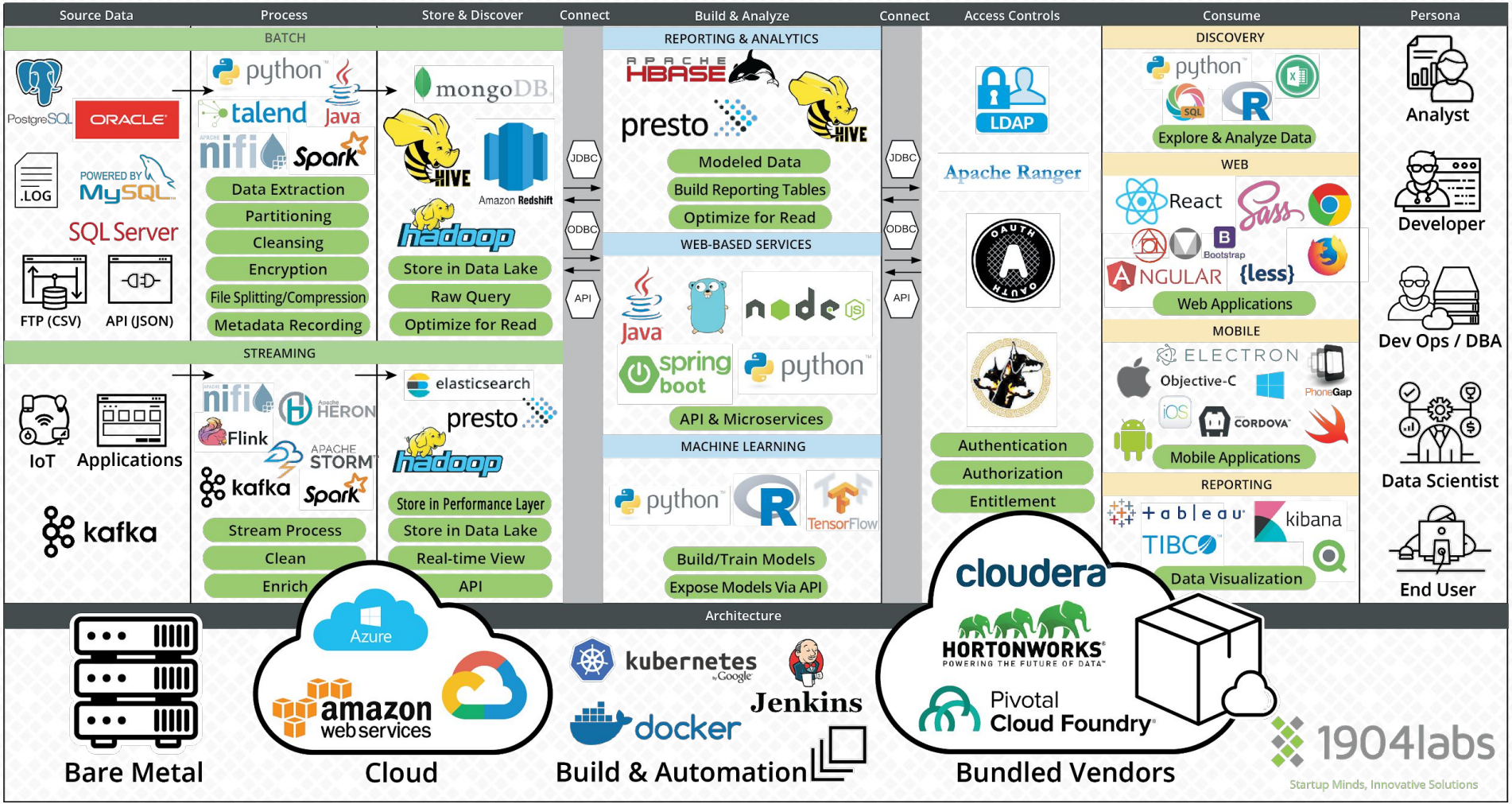
Discussion Topics

- Introductions
- Background - Chris
- Distributed systems - Chris
- Day in the life - Tim
- Demo - Kit

1904labs

Why Data Engineering?

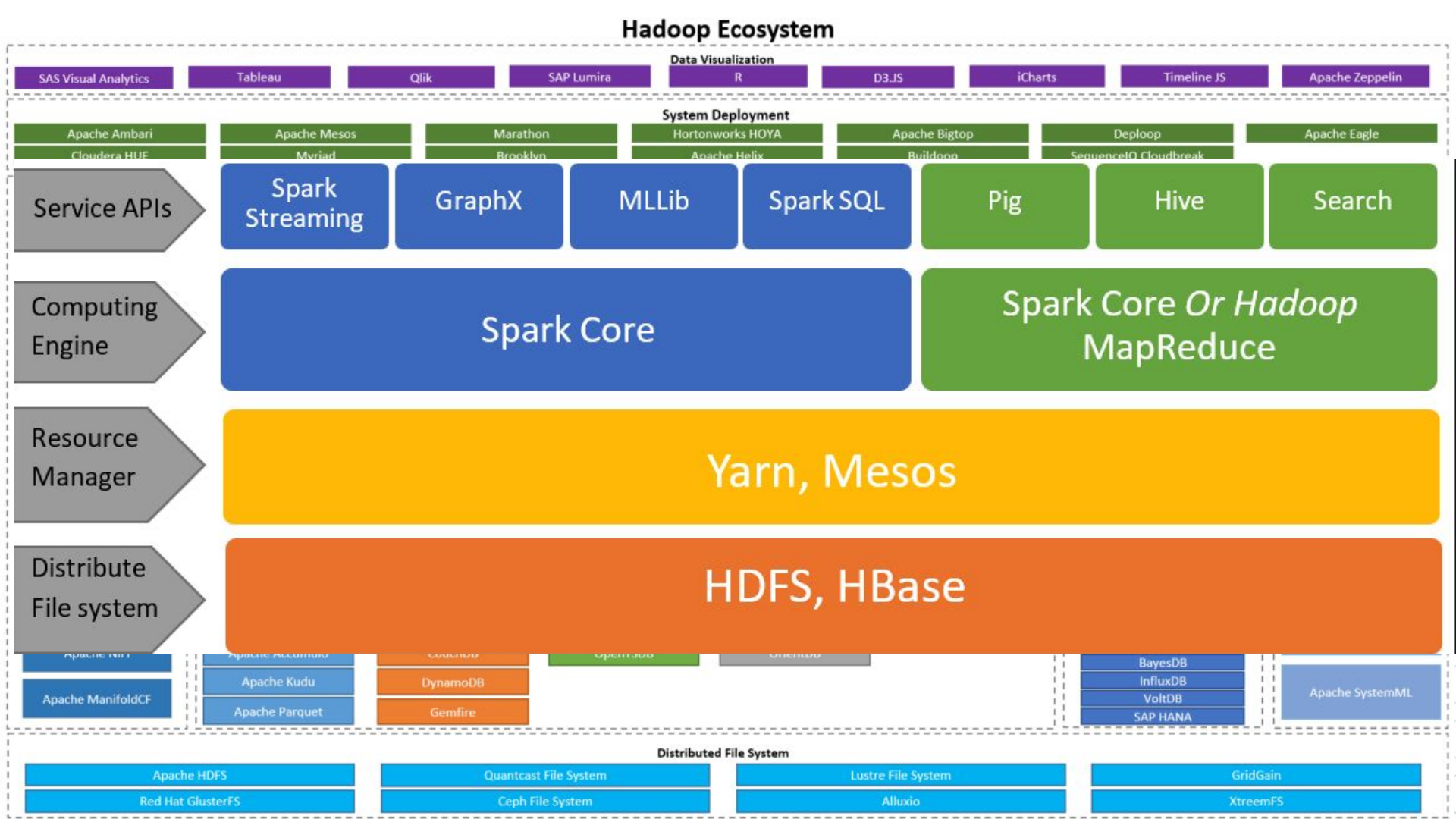
- It's not just about analytics.
- Treating data as an asset.
- Data science is built upon the work of good data engineering.
- The data we collect is changing. The way we store and access it must as well.
- Automate, automate, automate.


Source Data
Process
Store & Discover
Connect
Build & Analyze
Connect
Access Controls
Consume
Persona
BATCH
PostgreSQL
ORACLE
.LOG
POWERED BY MySQL
SQL Server
FTP (CSV)
API (JSON)
python
talend
Java
nifi
Spark
Data Extraction
Partitioning
Cleansing
Encryption
File Splitting/Compression
Metadata Recording
mongoDB.
HIVE
Amazon Redshift
hadoop
Store in Data Lake
Raw Query
Optimize for Read
JDBC
ODBC
API
REPORTING & ANALYTICS
APACHE HBASE
presto
HIVE
Modeled Data
Build Reporting Tables
Optimize for Read
WEB-BASED SERVICES
Java
node
spring boot
python
API & Microservices
MACHINE LEARNING
python
R
TensorFlow
Build/Train Models
Expose Models Via API
JDBC
ODBC
API
LDAP
Apache Ranger
OAUTH
Authentication
Authorization
Entitlement
DISCOVERY
python
SQL
R
Explore & Analyze Data
WEB
React
Sass
Bootstrap
ANGULAR
{less}
Web Applications
MOBILE
ELECTRON
Objective-C
PhoneGap
iOS
CORDOVA
Mobile Applications
REPORTING
+ableau
TIBCO
kibana
Data Visualization
Analyst
Developer
Dev Ops / DBA
Data Scientist
End User
STREAMING
IoT
Applications
kafka
nifi
Apache HERON
Flink
APACHE STORM
kafka
Spark
Stream Process
Clean
Enrich
elasticsearch
presto
hadoop
Store in Performance Layer
Store in Data Lake
Real-time View
API
Architecture
Azure
amazon web services
Bare Metal
Cloud
kubernetes
Jenkins
docker
Build & Automation
cloudera
HORTONWORKS
POWERING THE FUTURE OF DATA
Pivotal Cloud Foundry
Bundled Vendors
1904labs
Startup Minds, Innovative Solutions

Distributed Systems

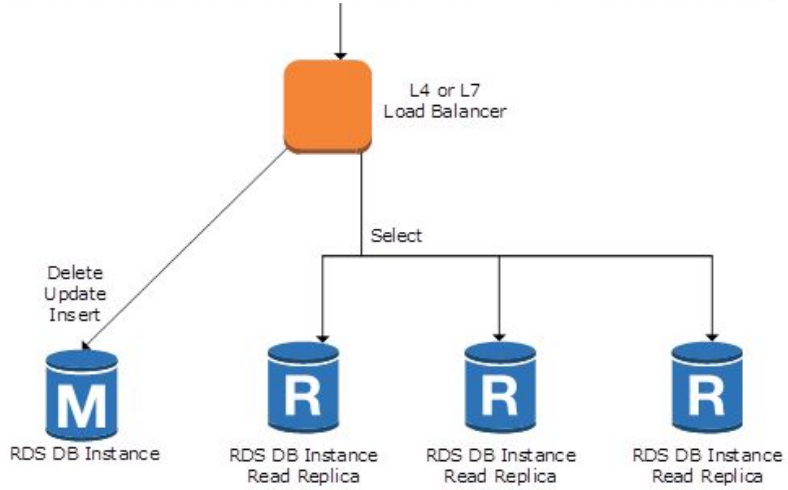
What does it mean?

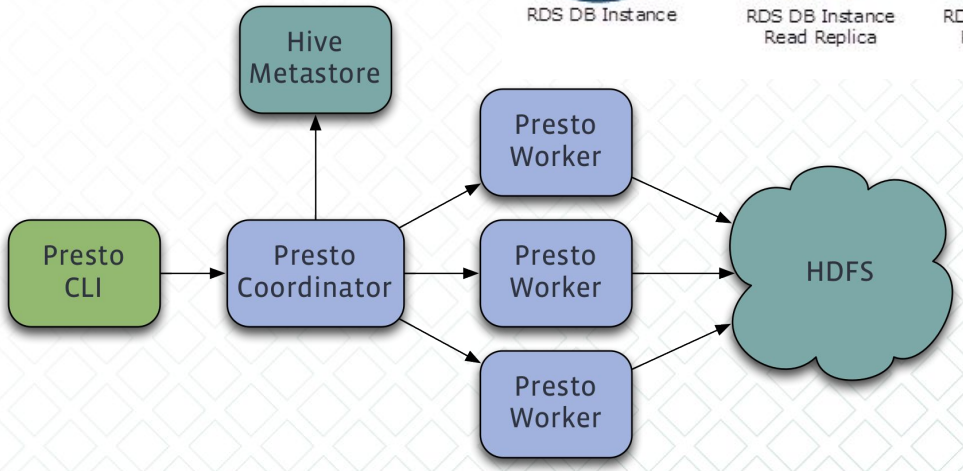
A distributed system is a system whose components are located on different networked computers, which communicate and coordinate their actions by passing messages to one another.


Hadoop Ecosystem
Data Visualization
SAS Visual Analytics
Tableau
Qlik
SAP Lumira
R
D3.JS
iCharts
Timeline JS
Apache Zeppelin
System Deployment
Apache Ambari
Cloudera HUE
Apache Mesos
Myriad
Marathon
Brooklyn
Hortonworks HOYA
Apache Helix
Apache Bigtop
Buildoop
Deploop
SequenceIQ Cloudbreak
Apache Eagle
Service APIs
Spark Streaming
GraphX
MLLib
Spark SQL
Pig
Hive
Search
Computing Engine
Spark Core
Spark Core Or Hadoop MapReduce
Resource Manager
Yarn, Mesos
Distribute File system
HDFS, HBase
Apache NiFi
Apache ManifoldCF
Apache Accumulo
Apache Kudu
Apache Parquet
CouchDB
DynamoDB
Gemfire
OpenTSDB
OrientDB
BayesDB
InfluxDB
VoltDB
SAP HANA
Apache SystemML
Distributed File System
Apache HDFS
Red Hat GlusterFS
Quantcast File System
Ceph File System
Lustre File System
Alluxio
GridGain
XtreemFS

Distributed Compute

1. Overview
2. Scaling
3. What does it mean for you?



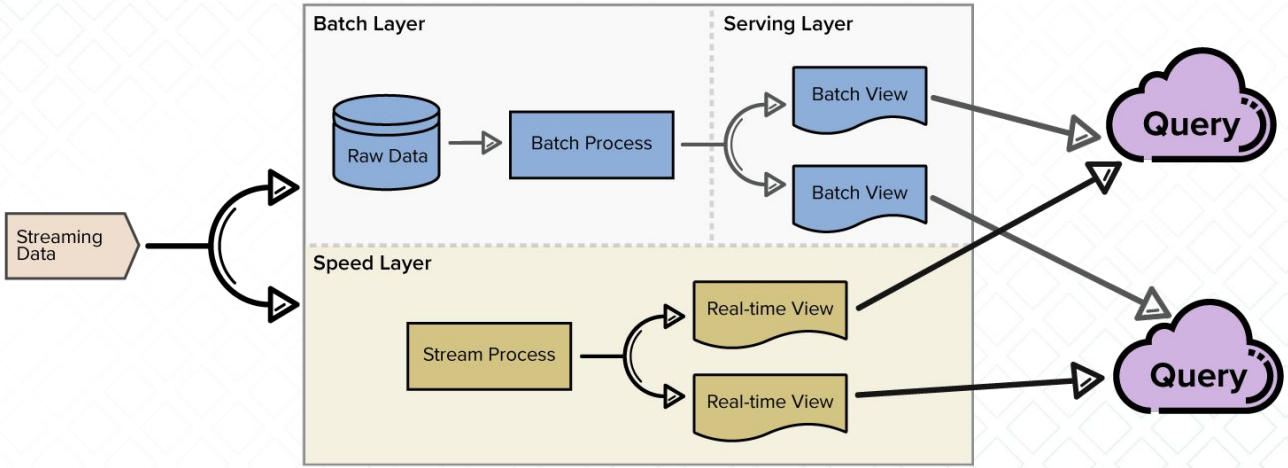


1904labs

Data Architecture - Lambda & Kappa

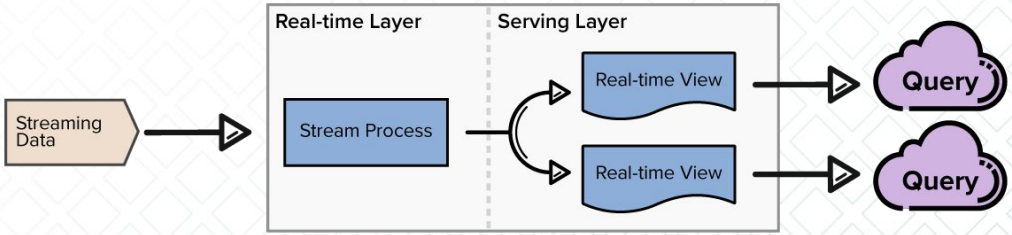
Lambda: Batch and Speed Layer

- Stores raw data
- Computes tables in batch
- Has streaming component
- Performs data modeling
- Exposes end tables to users



Kappa: Speed Layer

- All data as a stream
- In stream processing
- Real-time views

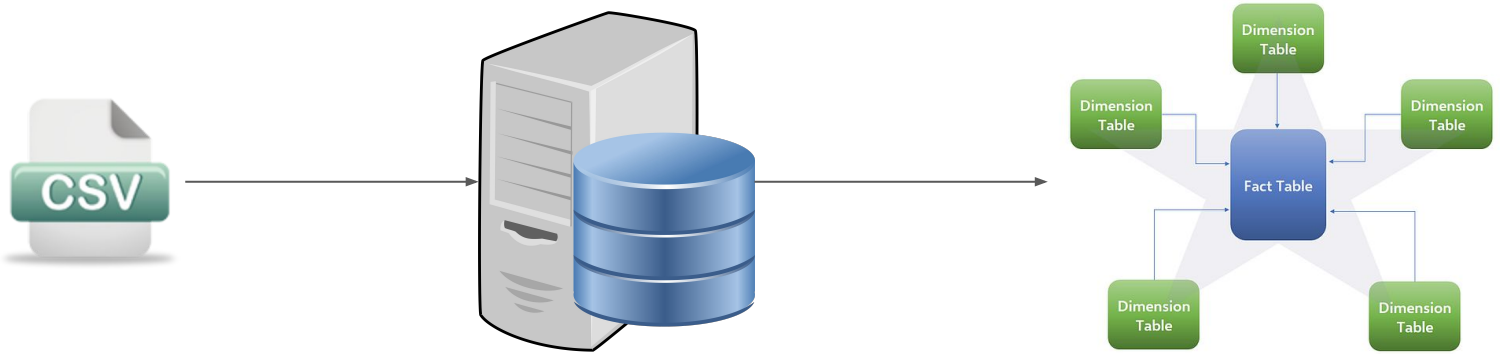


1904labs

A Day In The Life

What's not modern about this architecture?

- Cron job runs once a day to check for presence of a file on an FTP server
- CSV file is downloaded and loaded into an RDBMS star schema



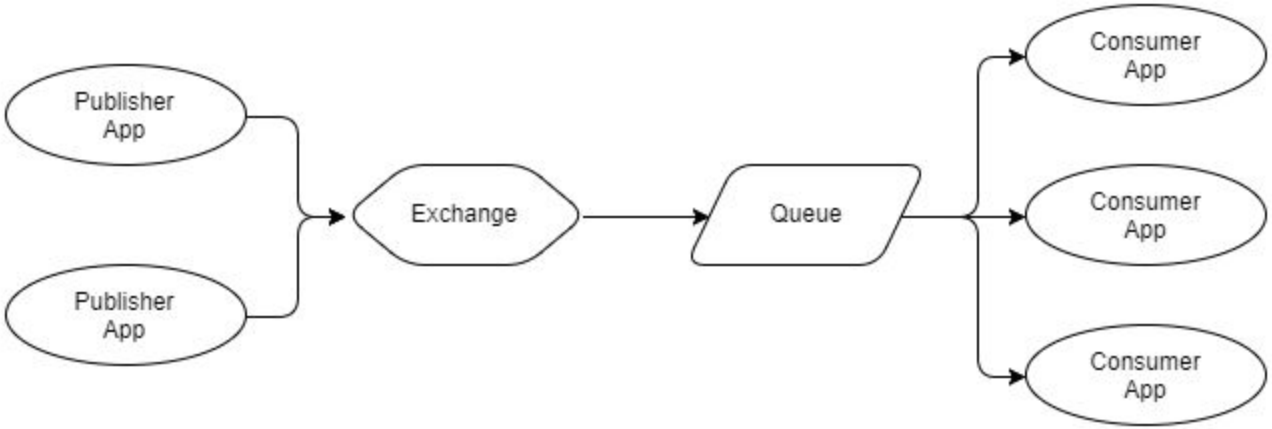
Message Queues

- Publishers
- Exchanges
- Queues
- Consumers

- Messages are removed from queues once a consumer reads them

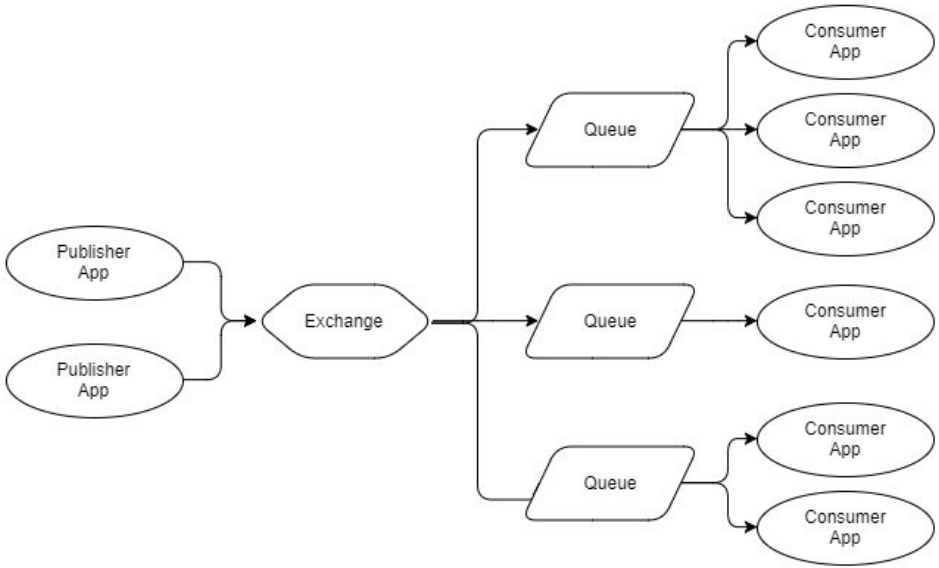
Scaling Up Message Queues

What if a single consumer can’t keep up with the volume?



Scaling Them Up Even Further...

What happens if each message needs to go to more than one set of consumers?



Distributed Logs

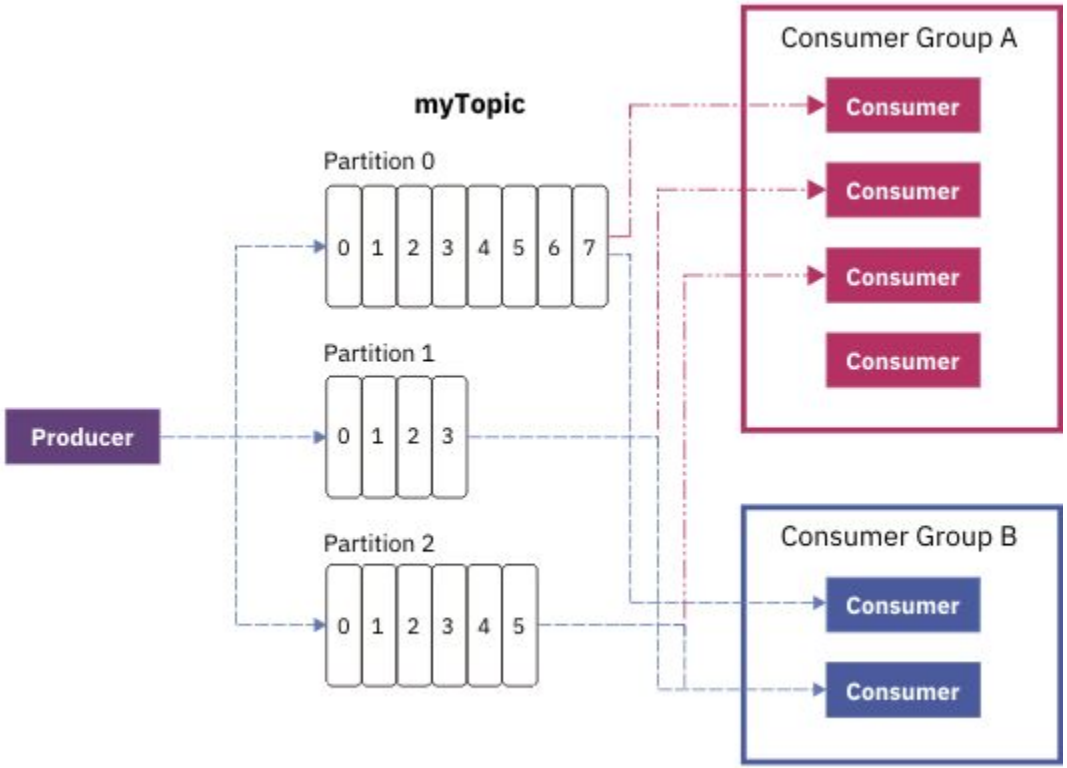
- Topic - a named stream of messages
- Partition - an ordered list of message. A topic is made up of a set of partitions
- Producer - a process that publishes messages to a topic
- Consumer - a process that consumes messages from a topic
- Consumer group - a named group of consumers who together consume messages from a set of topics

Topics can be consumed by multiple consumer groups independently.

Each consumer group is responsible for tracking its own position in message processing

Messages stay in the log for the defined retention period (are not removed once all consumer groups have processed them)

Distributed Logs



Data Formats

Apache Avro

- Row oriented
- Fast at writes
- Rich support for schema evolution

```
{
   "type" : "record",
   "namespace" : "STL Big Data",
   "name" : "Inventory",
   "fields" : [
     { "name" : "date"  , "type" : "date" },
     { "name" : "price" , "type" : "int" },
     { "name" : "size"  , "type" : "int" }
   ]
}
```

date	price	size
2011-01-20	10.1	10
2011-01-21	10.3	20
2011-01-22	10.5	40
2011-01-23	10.4	5
2011-01-24	11.2	55
2011-01-25	11.4	66
...	...	...
2013-03-31	17.3	100

Apache Avro

- Row oriented
- Fast at writes
- Rich support for schema evolution

```
{
   "type" : "record",
   "namespace" : "STL Big Data",
   "name" : "Inventory",
   "fields" : [
     { "name" : "date"  , "type" : "date" },
     { "name" : "price" , "type" : "int" },
     { "name" : "size"  , "type" : "int" },
     { "name" : "qty", "type" : [ "null", "int" ], "default" : null }
   ]
}
```

date	price	size
2011-01-20	10.1	10
2011-01-21	10.3	20
2011-01-22	10.5	40
2011-01-23	10.4	5
2011-01-24	11.2	55
2011-01-25	11.4	66
…	…	…
2013-03-31	17.3	100

Column Oriented Storage

- Instead of storing all columns for a row together on disk, store each column together on disk
- Enables higher compression ratios
- What kind of queries will you be executing against the data?

date	price	size
2011-01-20	10.1	10
2011-01-21	10.3	20
2011-01-22	10.5	40
2011-01-23	10.4	5
2011-01-24	11.2	55
2011-01-25	11.4	66
...	...	...
2013-03-31	17.3	100

date	price	size
2011-01-20	10.1	10
2011-01-21	10.3	20
2011-01-22	10.5	40
2011-01-23	10.4	5
2011-01-24	11.2	55
2011-01-25	11.4	66
...	...	...
2013-03-31	17.3	100

Apache Parquet

- Column oriented
- Enables fast reading
- Compresses well
- Commonly used with with Impala

Apache ORC

- Column oriented
- Enables fast reading
- Compresses extremely well
- Commonly used with Hive and Presto

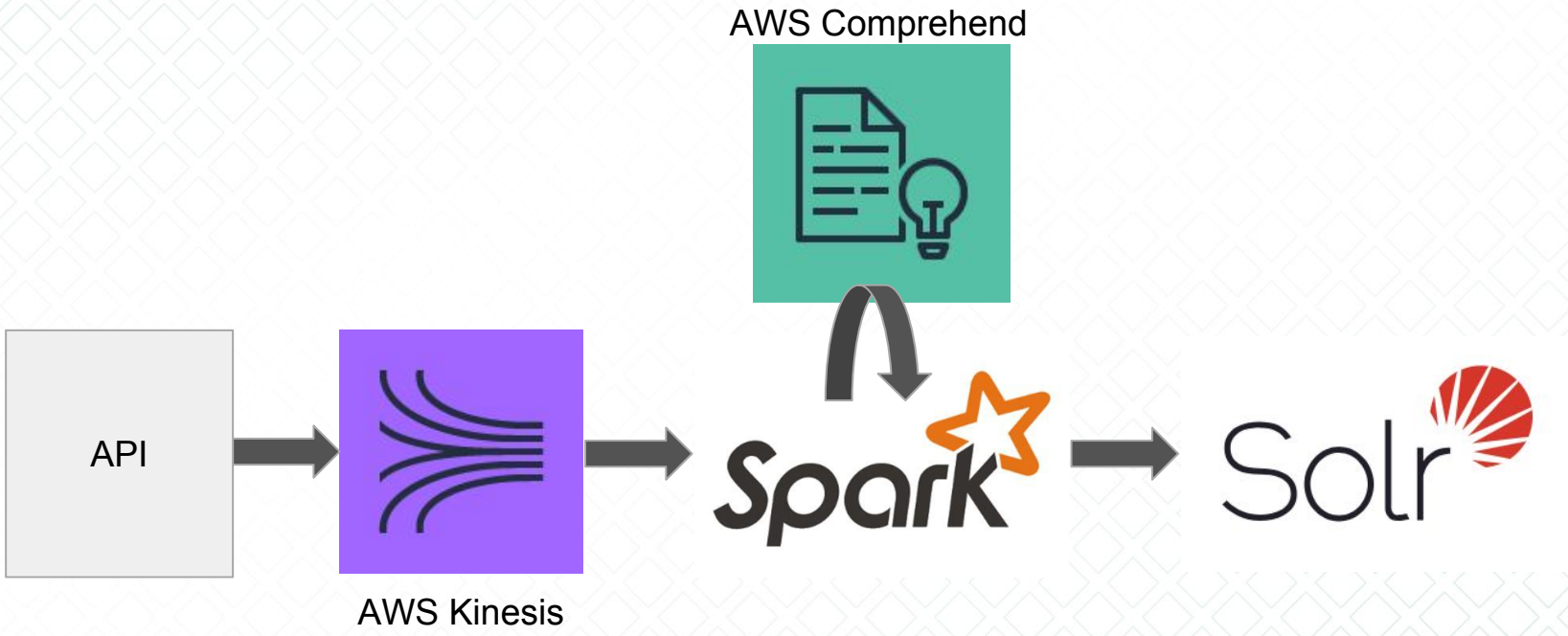
Compression Comparison

Format	Size
CSV	2.5 GB
Avro	1.2 GB
Parquet	817 MB
ORC	796 MB

Source: Netflix Prize Data https://www.kaggle.com/netflix-inc/netflix-prize-data

Demo

https://github.com/kitmenke/spark-hello-world


AWS Comprehend
API
AWS Kinesis
Spark
Solr

Appendix

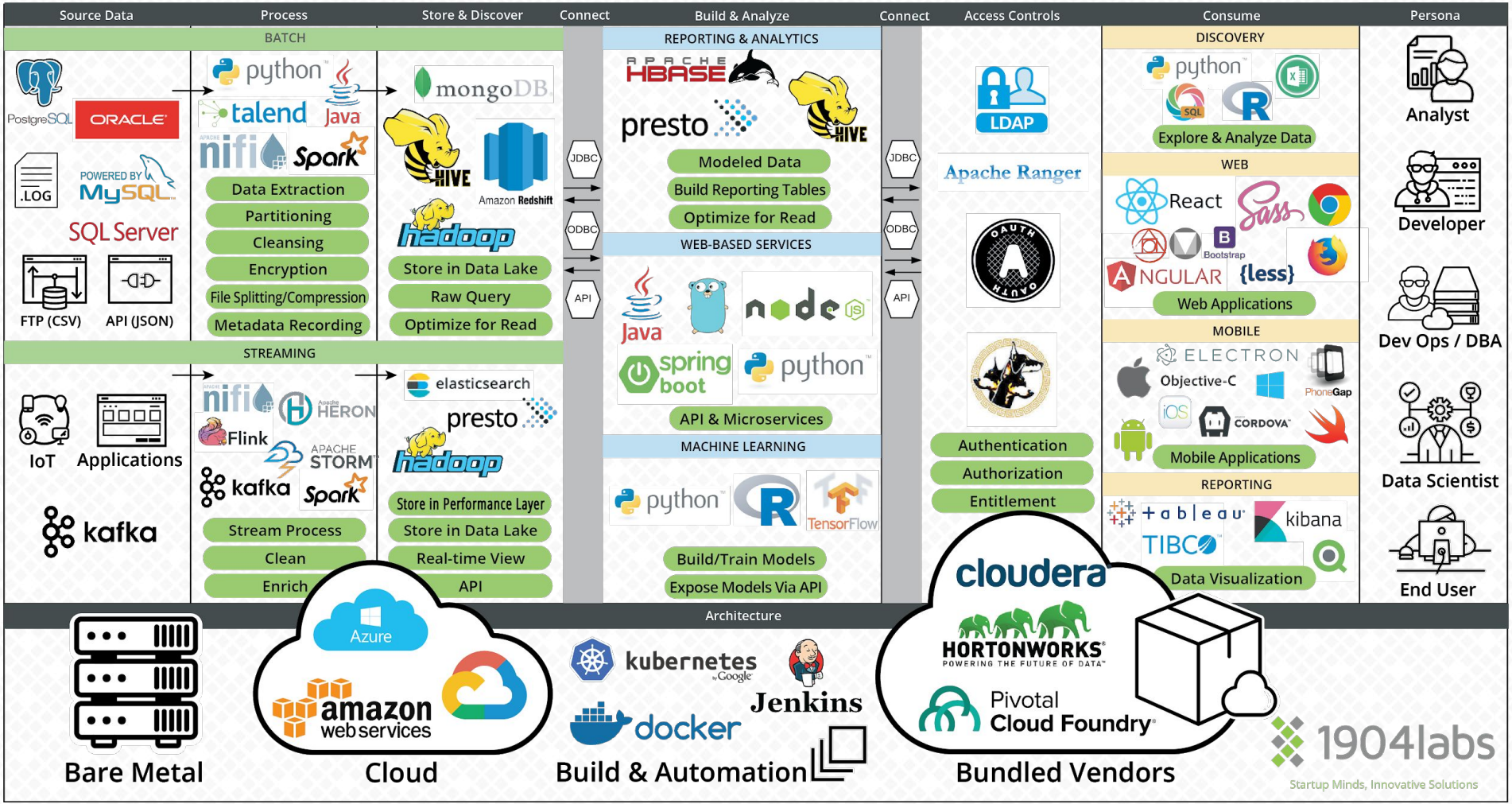
Building the Right Thing the Right Way

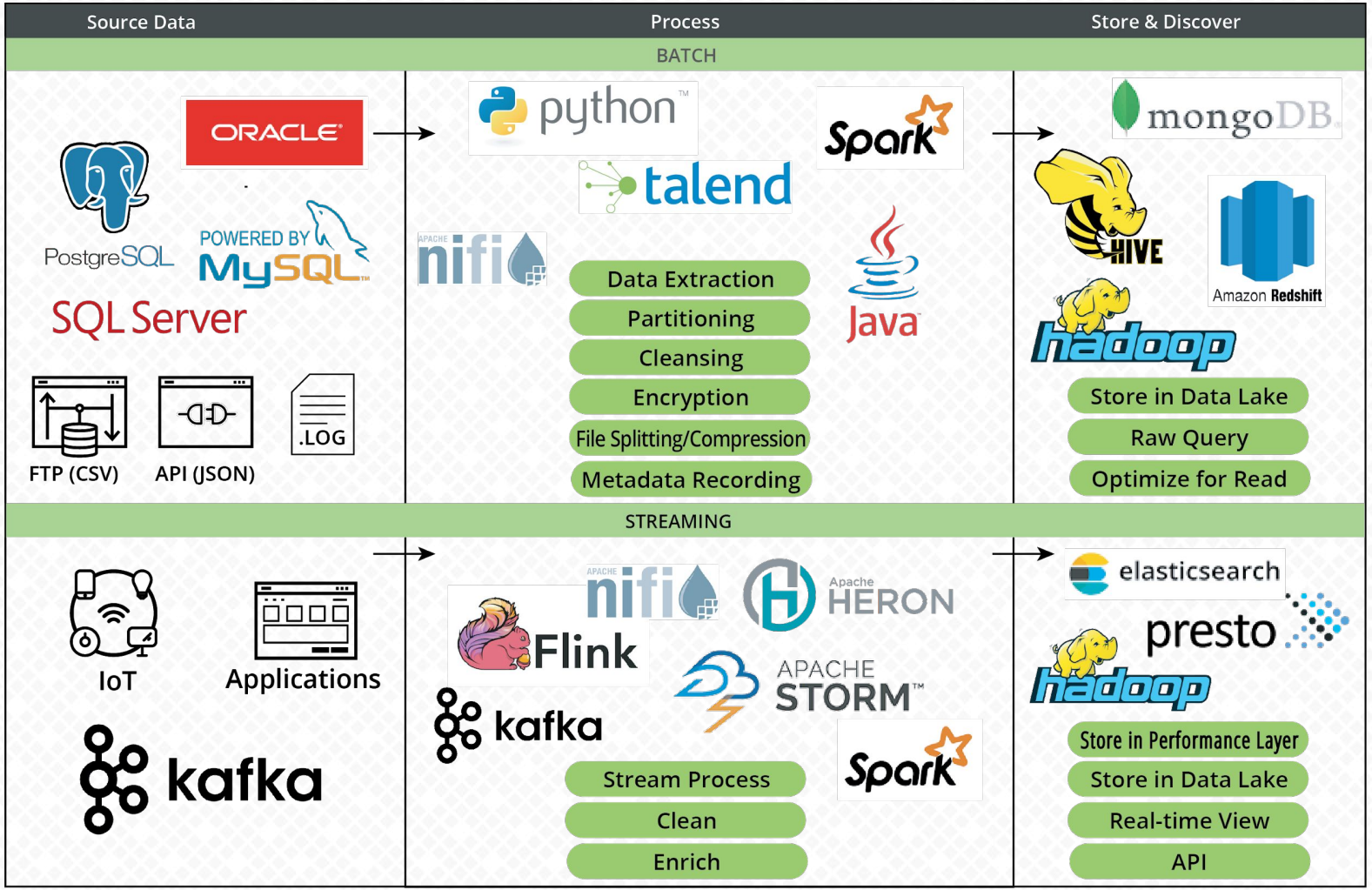
We listen to and understand our clients' most complex business challenges and bring an innovative approach to building modern technology solutions.

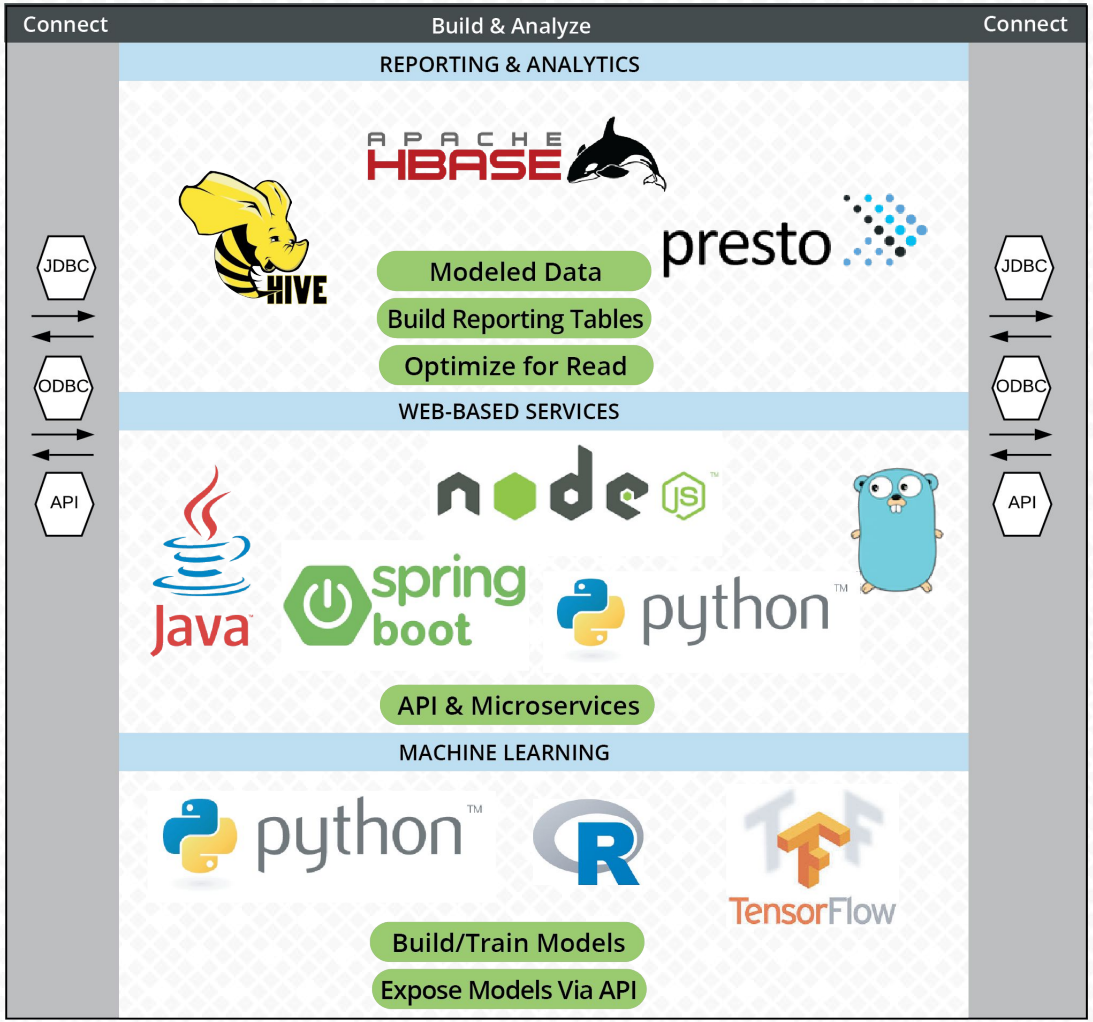
1904labs

Best in class modern architecture

... At 1904labs, we have a passion for open source technologies, strive to build cloud first applications, and are motivated by our desire to transform businesses into data-driven enterprises.


Source Data
Process
Store & Discover
Connect
Build & Analyze
Connect
Access Controls
Consume
Persona
BATCH
PostgreSQL
ORACLE
.LOG
POWERED BY MySQL
SQL Server
FTP (CSV)
API (JSON)
python
talend
Java
nifi
Spark
Data Extraction
Partitioning
Cleansing
Encryption
File Splitting/Compression
Metadata Recording
mongoDB.
HIVE
Amazon Redshift
hadoop
Store in Data Lake
Raw Query
Optimize for Read
JDBC
ODBC
API
REPORTING & ANALYTICS
APACHE HBASE
presto
HIVE
Modeled Data
Build Reporting Tables
Optimize for Read
WEB-BASED SERVICES
Java
node
spring boot
python
API & Microservices
MACHINE LEARNING
python
R
TensorFlow
Build/Train Models
Expose Models Via API
JDBC
ODBC
API
LDAP
Apache Ranger
OAUTH
Authentication
Authorization
Entitlement
DISCOVERY
python
SQL
R
Explore & Analyze Data
WEB
React
Sass
Bootstrap
ANGULAR
{less}
Web Applications
MOBILE
ELECTRON
Objective-C
PhoneGap
iOS
CORDOVA
Mobile Applications
REPORTING
+ableau
TIBCO
kibana
Data Visualization
Analyst
Developer
Dev Ops / DBA
Data Scientist
End User
STREAMING
IoT
Applications
kafka
nifi
Apache HERON
Flink
APACHE STORM
kafka
Spark
Stream Process
Clean
Enrich
elasticsearch
presto
hadoop
Store in Performance Layer
Store in Data Lake
Real-time View
API
Architecture
Azure
amazon web services
Bare Metal
Cloud
kubernetes
Jenkins
docker
Build & Automation
cloudera
HORTONWORKS
POWERING THE FUTURE OF DATA
Pivotal Cloud Foundry
Bundled Vendors
1904labs
Startup Minds, Innovative Solutions

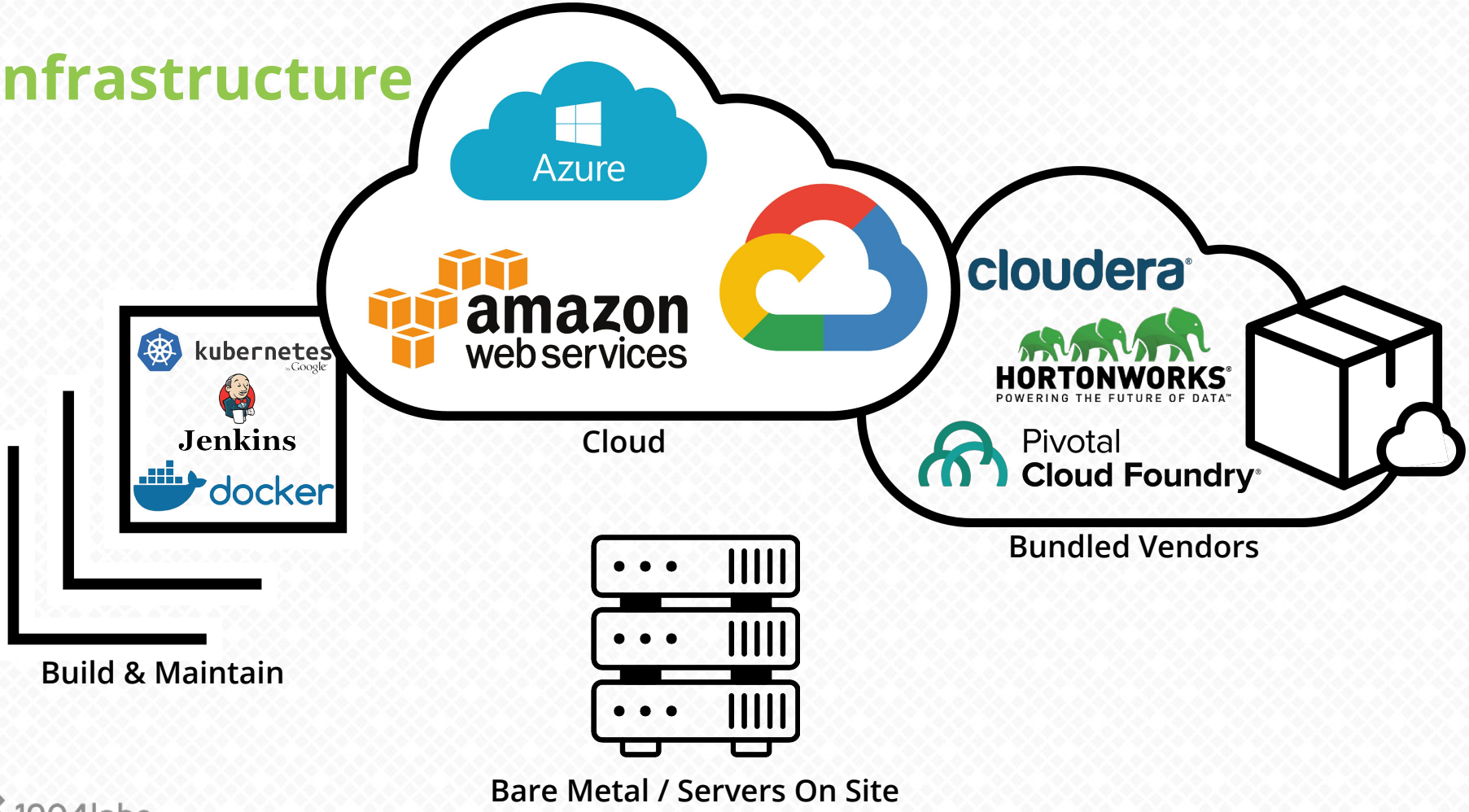

Source Data
Process
Store & Discover
BATCH
ORACLE
PostgreSQL
POWERED BY MySQL
SQL Server
FTP (CSV)
API (JSON)
.LOG
python
Spark
talend
nifi
Java
Data Extraction
Partitioning
Cleansing
Encryption
File Splitting/Compression
Metadata Recording
mongoDB
HIVE
Amazon Redshift
hadoop
Store in Data Lake
Raw Query
Optimize for Read
STREAMING
IoT
Applications
kafka
nifi
Apache HERON
Flink
APACHE STORM
kafka
Spark
Stream Process
Clean
Enrich
elasticsearch
presto
hadoop
Store in Performance Layer
Store in Data Lake
Real-time View
API


Connect
Build & Analyze
Connect
REPORTING & ANALYTICS
APACHE HBASE
HIVE
presto
Modeled Data
Build Reporting Tables
Optimize for Read
JDBC
ODBC
API
JDBC
ODBC
API
WEB-BASED SERVICES
node
Java
spring boot
python
API & Microservices
MACHINE LEARNING
python
R
TensorFlow
Build/Train Models
Expose Models Via API

1904labs


Access Controls
Consume
Persona
DISCOVERY
python
SQL
R
Explore & Analyze Data
WEB
React
ANGULAR
Post CSS
Material Design
Bootstrap
Sass
{less}
Web Applications
MOBILE
ELECTRON
Objective-C
iOS
CORDOVA
PhoneGap
Mobile Applications
REPORTING
+ableau
TIBCO
kibana
Data Visualization
LDAP
Apache Ranger
OAUTH
Authentication
Authorization
Entitlement
Analyst
Developer
Dev Ops / DBA
Data Scientist
End User

Infrastructure



Source Data - Data as it exists in your current company infrastructure

Process - Pulling source data while enriching, cleansing, and encrypting the data in transit.

Store & Discover - Inserting data into modern architecture. Raw data can be queried at this level.

Build & Analyze - The storage layer, data models, and web services are built and consumed by reporting services and end users.

Connect - Both external users & internal components will typically connect through JDBC, ODBC, and API.

Access Controls - Standard security layer used for authentication & authorization of all services.

Consume - The applications which connect to the underlying modern architecture.

Persona - All people involved with the modern architecture. Some will be international developers, while others are external.

Source Data

BATCH

PostgreSQL
ORACLE
.LOG
MySQL
SQL Server
FTP (CSV)
API (JSON)

STREAMING

IoT
Applications
kafka

Process

BATCH

python
talend
Java
nifi
Spark
Data Extraction
Partitioning
Cleansing
Encryption
File Splitting/Compression
Metadata Recording

STREAMING

nifi
Apache HERON
Flink
APACHE STORM
kafka
Spark
Stream Process
Clean
Enrich

Store & Discover

BATCH

mongoDB
HIVE
Amazon Redshift
hadoop
Store in Data Lake
Raw Query
Optimize for Read

STREAMING

elasticsearch
presto
hadoop
Store in Performance Layer
Store in Data Lake
Real-time View
API

Connect

JDBC
ODBC
API

Build & Analyze

REPORTING & ANALYTICS

APACHE HBASE
HIVE
presto
Modeled Data
Build Reporting Tables
Optimize for Read

WEB-BASED SERVICES

Java
node
spring boot
python
API & Microservices

MACHINE LEARNING

python
R
TensorFlow
Build/Train Models
Expose Models Via API

Connect

JDBC
ODBC
API

Access Controls

LDAP
Apache Ranger
OAUTH
Authentication
Authorization
Entitlement

Consume

DISCOVERY

python
SQL
R
Explore & Analyze Data

WEB

React
Sass
Bootstrap
ANGULAR
{less}
Web Applications

MOBILE

ELECTRON
Objective-C
PhoneGap
iOS
CORDOVA
Mobile Applications

REPORTING

+ableau
kibana
TIBCO
Data Visualization

Persona

Analyst
Developer
Dev Ops / DBA
Data Scientist
End User

Architecture

Bare Metal

Cloud
Azure
amazon web services

Build & Automation
kubernetes
Jenkins
docker

Bundled Vendors
cloudera
HORTONWORKS
POWERING THE FUTURE OF DATA
Pivotal Cloud Foundry

1904labs
Startup Minds, Innovative Solutions

Open-Source Adoption in the Industry

These open source tools are already used broadly across industries today

1904labs' primary focus is to help our clients design, build and operate world class and modern data, development operations and decision science capabilities…

…Utilizing best of breed open source and/or commercial open source tools and platforms.