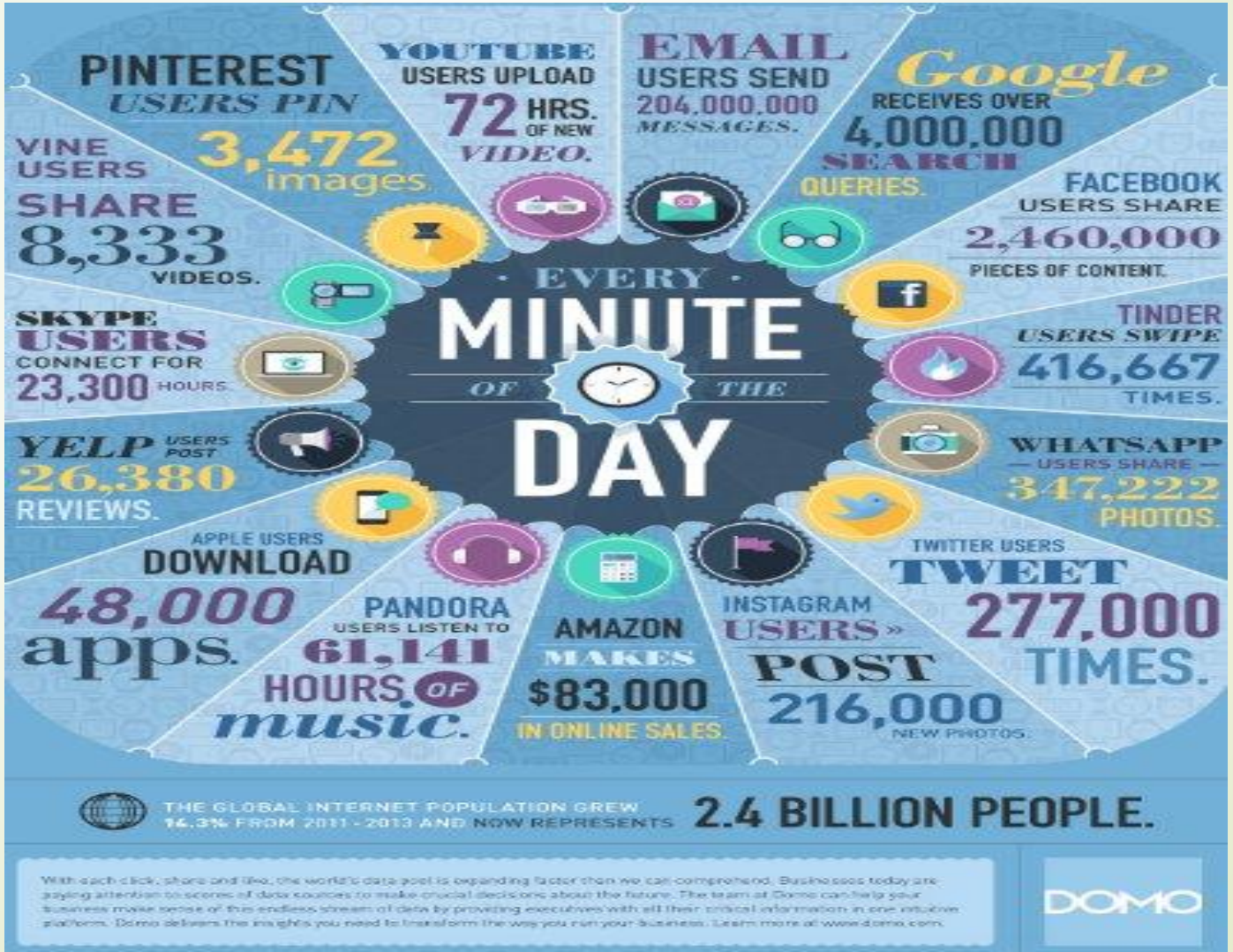
Data Processing with Apache Spark

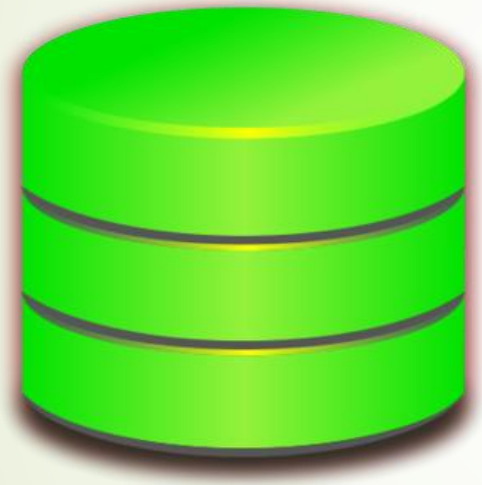
Harsha Puthalapattu

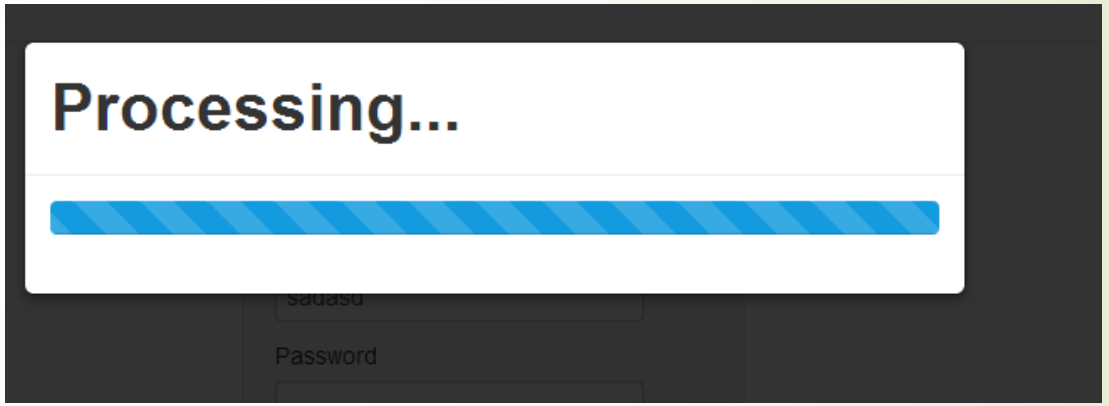
@harshappt

Big Data


PINTEREST
USERS PIN
3,472 images
YOUTUBE
USERS UPLOAD
72 HRS. OF NEW
VIDEO.
EMAIL
USERS SEND
204,000,000
MESSAGES.
Google
RECEIVES OVER
4,000,000
SEARCH
QUERIES.
VINE
USERS
SHARE
8,333
VIDEOS.
FACEBOOK
USERS SHARE
2,460,000
PIECES OF CONTENT.
EVERY
MINUTE
OF THE
DAY
SKYPE
USERS
CONNECT FOR
23,300 HOURS.
TINDER
USERS SWIPE
416,667
TIMES.
YELP USERS POST
26,380
REVIEWS.
WHATSAPP
USERS SHARE
347,222
PHOTOS.
APPLE USERS
DOWNLOAD
48,000
apps.
PANDORA
USERS LISTEN TO
61,141
HOURS OF
music.
AMAZON
MAKES
$83,000
IN ONLINE SALES.
INSTAGRAM
USERS
POST
216,000
NEW PHOTOS.
TWITTER USERS
TWEET
277,000
TIMES.
THE GLOBAL INTERNET POPULATION GREW
14.3% FROM 2011 - 2013 AND NOW REPRESENTS
2.4 BILLION PEOPLE.
DOMO

Big Data Problem





Big Data Problem

- Commodity hardware
- Cheaper
- Easy to add as data grows

- Non Reliable
- Performance Issues

- How to handle data serialization?
- How to ensure data integrity?
- How to recover failures?
- How to store and retrieve data?

Big Data Problem

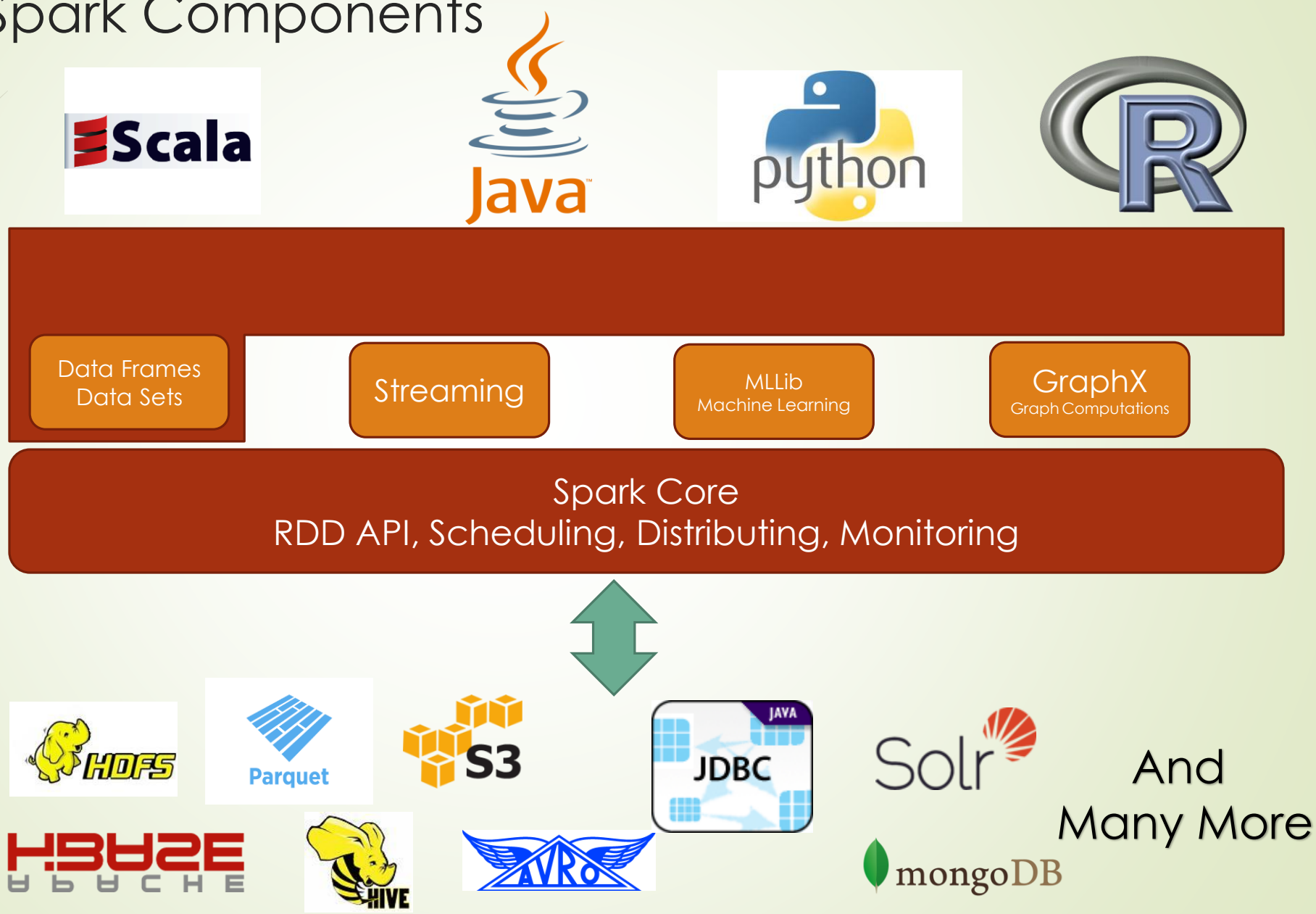
IMPALA

- Map Reduce is Slow in large jobs
- Specialized processing requires specialized tools
- Data Engineers and Analysts need to learn multiple frameworks and tools

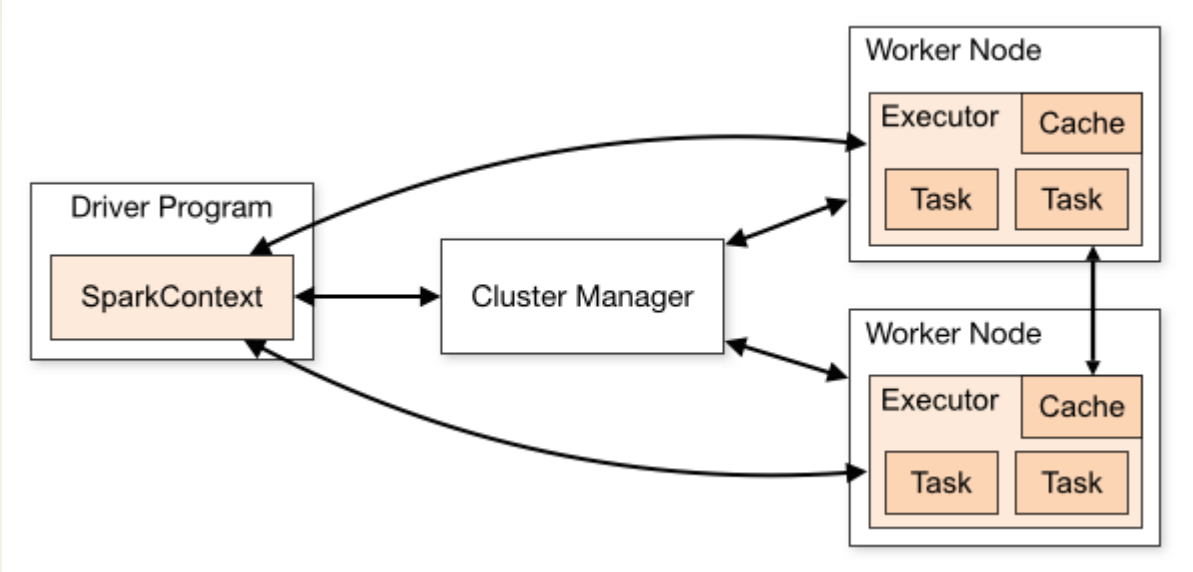
Spark - Fast and General purpose Cluster Computing System

- AMP Labs, UC Berkeley
- Paper published in 2010
- Apache top level project in 2014
- Evolved as ecosystem
- Primarily provides – Scheduling, Monitoring and Distributing Capabilities
- Unified API

Spark Components



Spark Execution Architecture



- Driver is the process running the main() function of the application and creating the SparkContext
- Cluster Manager is an external service for acquiring resources on the cluster (e.g. standalone manager, Mesos, YARN)
- Worker Node is any node that can run application code in the cluster
- Executor is a process launched for an application on a worker node, that runs tasks and keeps data in memory or disk storage across them. Each application has its own executors

Data Abstraction

RDD

- An immutable collection elements partitioned across the nodes of the cluster that can be operated on in parallel.

Dataset/Dataframe

- A Dataset is a strongly typed collection of domain-specific objects that can be transformed in parallel using functional or relational operations. Each Dataset also has an untyped view called a DataFrame, which is a Dataset of Row

Demo

RDD and DataSet