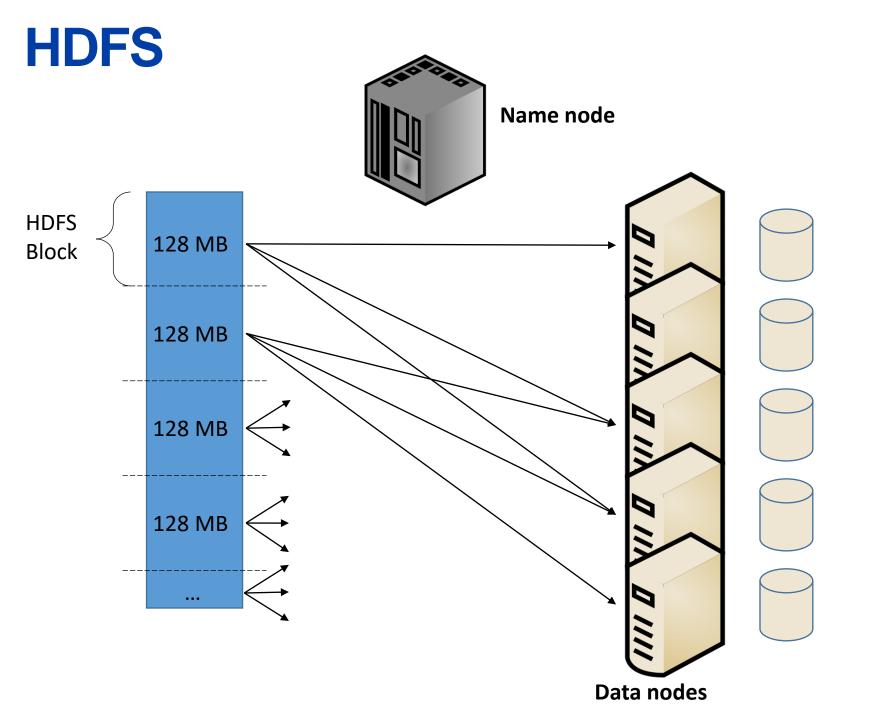
#### Introduction to Big-data Management

Review and next steps



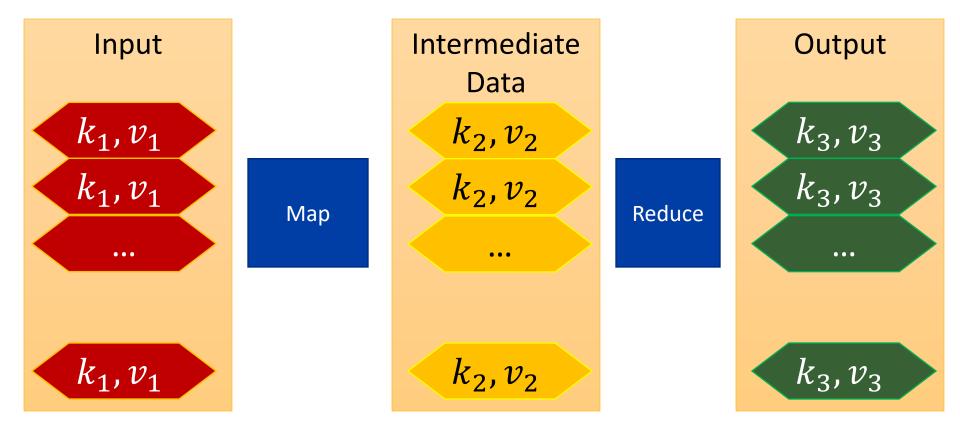
### What We Covered

- Storage (HDFS)
- Query processing (MapReduce, RDD, Hyracks)
- Higher-level data flow engines (Pig, SparkSQL)
- Storage formats (row, column, Parquet, LSM indexing)
- Document databases (MongoDB)
- Machine learning (MLlib)



## Logical View of MapReduce

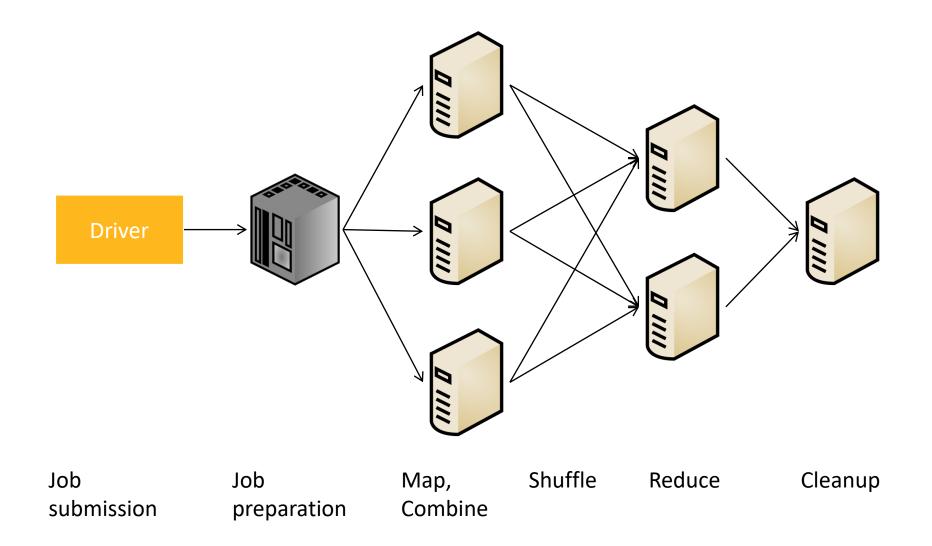
 During MapReduce, the input and output are considered a set of key-value pairs (k, v)



#### Map and Reduce Functions

- Map Function
  - Maps a single input record to a set (possibly empty) of intermediate records
  - Map:  $\langle k_1, v_1 \rangle \rightarrow \{\langle k_2, v_2 \rangle\}$
- Combine Function
  - Combine:  $\langle k_2, \{v_2\} \rangle \rightarrow \{\langle k_2, v_2 \rangle\}$
- Reduce Function
  - Reduces a set of intermediate records with the same key to a set (possibly empty) of output records
  - Reduce:  $\langle k_2, \{v_2\} \rangle \rightarrow \{\langle k_3, v_3 \rangle\}$

#### **Job Execution Overview**



#### **Resilient Distributed Dataset (RDD)**

- RDD is a pointer to a distributed dataset
- Stores information about how to compute the data rather than where the data is
- Transformation: Converts an RDD to another RDD
- Action: Returns an answer of an operation over an RDD
- Narrow Vs wide dependencies
- How RDD operations work

## **SparkSQL**

#### Dataframe (SparkSQL)

- Lazy execution
- Spark is aware of the data model
- Spark is aware of the query logic
- Can optimize the query

#### <u>RDD</u>

- Lazy execution
- The data model is hidden from Spark
- The transformations and actions are black boxes
- Cannot optimize the query

#### MLlib

- Main components of MLlib
  - Transformers, e.g., feature extraction
  - Estimator, e.g., clustering or regression
  - Evaluator, e.g., precision and recall calculation
  - Validator, e.g., k-fold cross validation
- Pipeline: Transformation(s) + Estimator

## **Big Spatial Data**

- How to customize Spark for a specific domain, i.e., spatial data
- Support various file formats other than the regular text files
- Build complex query pipelines such as spatial join and visualization
- Combine spatial operations with regular Spark operations

## **Storage formats**

- Difference between row and column formats
  - How attributes map to disk
  - Major applications for each of them
- Parquet files
  - A column store file format
  - Handles nesting and replication
  - Schema → Maximum definition and repetition level
  - Record → Definition and repetition level for each attribute
  - Do not forget to add null (non-existent) attributes

#### **Document databases**

- How a document database compares to a relational database (RDBMS)
  - Normalization (nesting and repetition)
  - ACID compliance
- How MongoDB compares attributes
- Log-structured-merge (LSM) tree for big data indexing

## Did we cover everything?



CELEBRATING 30 YEARS Marlan and Rosemary Bourns College of Engineering

#### **Big Data Landscape**

DATA	& AI	LANDSCAF	PE 2020
------	------	----------	---------

INFRASTRUCTURE	ANALYTICS & MACHINE INTELLIGENCE	APPLICATIONS – ENTERPRISE –				
STORAGE HADOOP CLOUDTRA Construction Con	BIPLATFORMS USUALIZATION DATA ANALYST PLATFORMS With the biscu init from the state of the state	SALES MARKETING- B28 MARKETING- B28 Product Actional Control C				
NoSQL DATABASES NewSQL DATABASE	DATA SCIENCE NOTEBOOKS Colob Subscience Colob Subscience Sub	LEGAL REGTECH & RANNE RIANCE Anaplan zuron URAN & RPA COMPUNITY &				
SCYLLA: @Amongoide @analysis and analysis an	COMPUTER VISION     COMPUTER VISION     COMPUTER VISION     COMPUTER VISION     COMPUTER VISION     COMPUTER VISION     Computer Contract     Compute	ADVERTISING EDUCATION REAL ESTATE SADVERTISING EDUCATION REAL ESTATE Criteol. LASterior Criteol. LASter				
MGMT / MONITORING MEMET /	SEARCH LIGG ANALYTICS Splunk: ↓ Splunk: ↓ Spluk: ↓ Splunk: ↓ Splunk: ↓ Spluk:	HEALTHCARE I Hatton X <sup>2</sup> (2003) <sup>1</sup> / <sub>2</sub> Mittarda <sup>1</sup> / <sub>2</sub> Hatton X <sup>2</sup> (2004) <sup>1</sup> / <sub>2</sub> Mittarda <sup>1</sup> / <sub></sub>				

#### **OPEN SOURCE** FRAMEWORKS QUERY / DATA FLOW -DATA ACCESS & DATABASES ORCHESTRATION STREAMING & STAT TOOLS & ALOPS AI / MACHINE LEARNING / DEEP LEARNING SEARCH LOGGING & MONITORING VISUALIZATION COLLABORATION 7 SECURITY & PIPELINES MESSAGING LANGUAGES & INFRA PM SODIK SOL **\_\_\_\_** - TANK Association in the Caffe antiner Soli<sup>®</sup> e ost raro th CO Superset Apache Range 🕈 Tenso 3 🗶 mif 📕 kibana Beaker talend 谊 R 2. 0matpl%tlib KNOX Arrest. sriak. FeatureFu Magan 📕 Scala 🛛 👔 Numby uaytei & SENTRY 👆 logstash Rant æ @Kabef ow DVC Sentry Ap Lamic 2 Diapri 🔹 Sphinx 🛛 💼 Zeppelin SLAMDATA Manana SciPy VELES andal Polyana (P OPrometheus 1 TensorBoard . 🙆 iifié Muerita 💙 uerit (0) Quant 💿 študio 🗿 🛛 julia accumulo COAP 🔕 HE101 **e**Flink OF VIOR DSSTNE fastati 0 🕄 Sank Thesis Second (antri) State of the local state of the ©Grafana ₩Vittion 📌 👹 snyk 🔶 😭 bokeh ANACONDA HELIX 20. 20 IN BENTOM. "mir 🤹 🚛 minded 📥 Red Hat W FMb

DATA SOURCES & APIs						DATA RESOURCES			
DATA MARKETPL		FINANCIAL & ECONOMIC DATA	AIR / SPACE / SEA	PEOPLE / ENTITIES	LOCATION INTELLIGENCE	OTHER	DATA SERVICES	INCUBATORS &	RESEARCH
& DISCOVERY		Bloomberg 🍪 THOMSON REUTERS 🕞   DOW JONES Quandl	🚫 окжызын лаевотсь 🛆 spire	Zzominio acxi@m. experian.		E OATA.GOV 🔊 MAN and			OpenAI facebook research
	Showficker Dete Deterge					IMAGENET LODO	Booz Allen Hemilton kaggle	DutaCamp A DataElite	
	🏇 EXPLORIUM		🌊 💥 🕹	Quantcast EASIS SAFEGRAPH	🗛 🛋 🕹 Radar	Berkeley DeepDrive	ElectritAi fractal XEXL	galvanize 🖗 🕬	
🚱 data.world	🗙 narrative		C. Synapective	Demyst melissa	CUEDIQ 🔊 OpenStreetMap	APDIO-SAR CRUX	DataKind innoPLCXUS	INSIG-II Cata Incubator	



## **Topics not Covered**

- Key-value stores
- Big graph analytics
- Visualization
- Streaming
- Coordination
- Cloud platforms

## **Key-value Stores**

- Provide a simple API to insert/delete/update/sear ch key-value pairs
- Records are indexed by key (typically a string)
- Internal structure is typically a Log-structuredmerge tree (LSM)
- Not generally suitable for large-scale analytics







## **Big Graph Analytics**

- Graphs are usually processed using a node-centric processing model
- Nodes and edges are both treated as first-class citizens
- Processing is normally iterative with a lot of iterations
   Ineo4j
   Ineo4j
   Ineo4j

## Visualization

- Sometimes called Business Intelligence
   (BI) <a href="mailto:versite">Rodeo</a> <a href="mailto:BeakerX">BeakerX</a>
- Focuses more on the end-user interface while producing nice graphs (e.g., bar charts and line graphs)
- Internally, the data is managed using the common big-data platforms but the systems are tuned to provide fast query response for ad-hoc queries

## Streaming

- Some applications need to process data in real-time with a very small latency
- Examples: Twitter search, IoT applications, and social network trends
- Works primarily off main memory
- Keeps only the latest records to ensure real-time response





## Coordination

- Most big-data systems are designed for shared-nothing large-scale analytics
- No coordination between machines is part of the design
- Coordination systems provide an easy way to coordinate the work in these distributed platforms, e.g., a catalog of information, work queue, and a global system status



### **Cloud Platforms**

- Maintaining your own cluster is costly
- It could be underutilized most of the time
- Cloud platforms allow you to rent virtual machines to do your work and dispose them after
- They are well-integrated with big data platforms (such as Hadoop and Spark) to give the best user experience
- All you need is an internet connection and a credit card



# What is next?

## What is next?

- Real big data is widely available
- Big data is like gold
- Only a few people know how to deal with it
- You're now one of them
- Applications
  - Keep your hands dirty

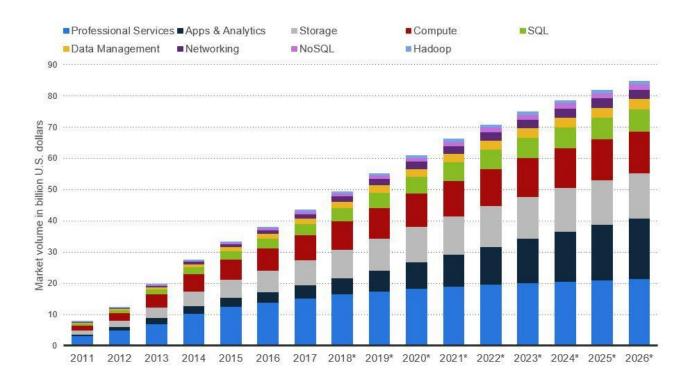


 Consider using the public cloud (e.g., AWS, Google Cloud, or Microsoft Azure)

#### **Job Market**

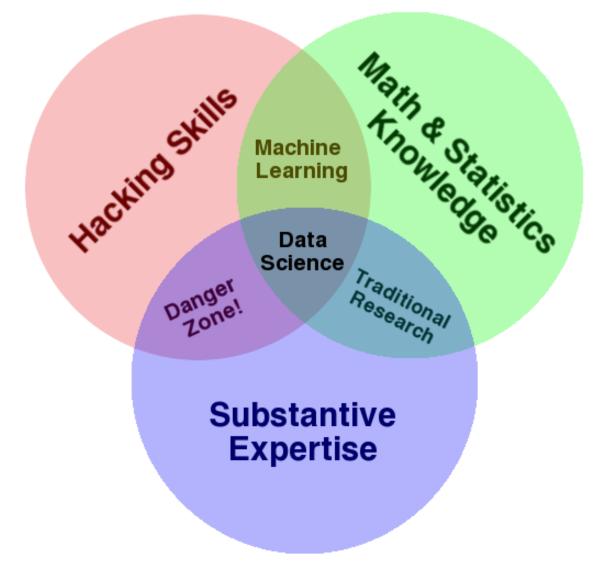
Big Data Market Worldwide Segment Revenue Forecast 2011-2026

Big Data Market Forecast Worldwide from 2011 to 2026, by segment (in billion U.S. dollars)



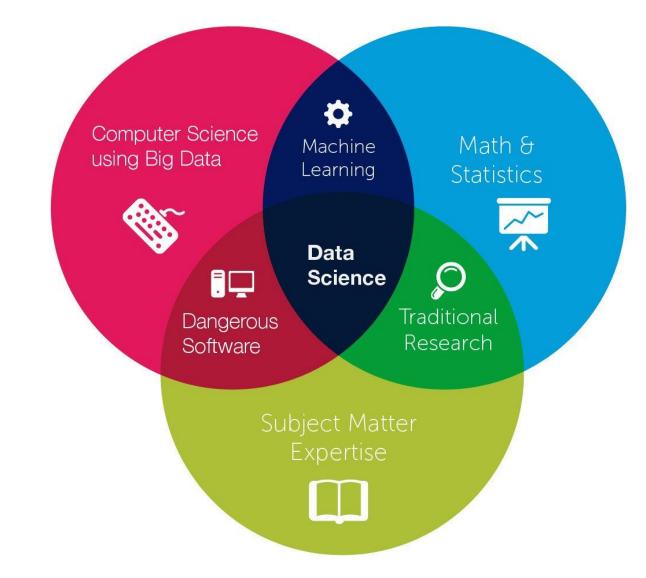


#### **Data Science**



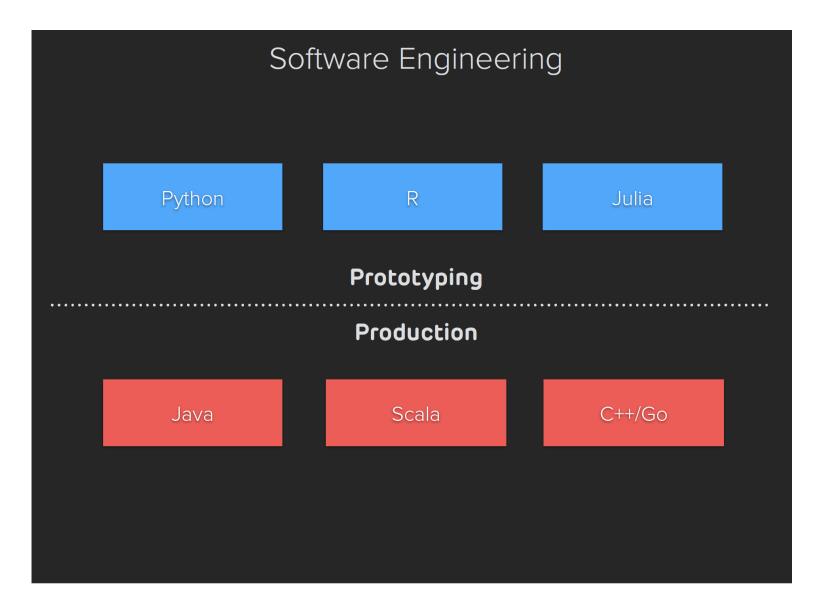
Credits: Drew Conway

#### **Data Science**

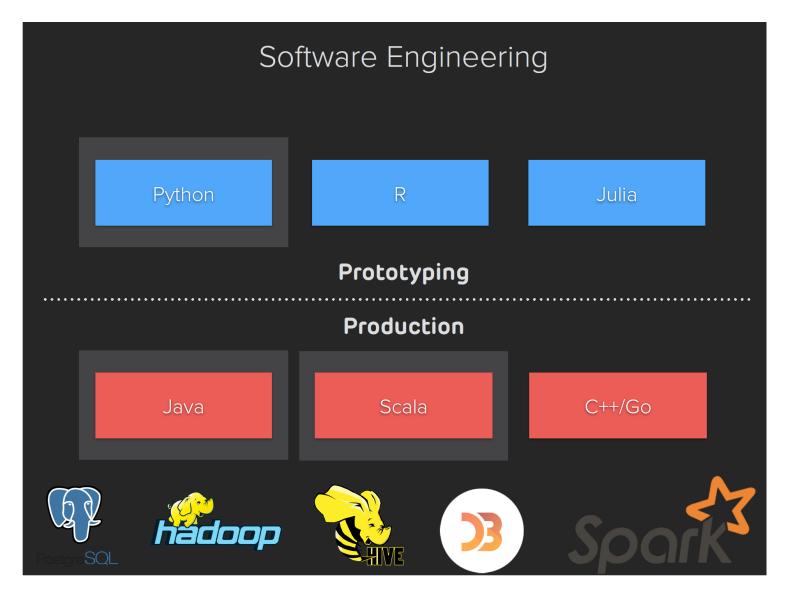


## **Next Steps**

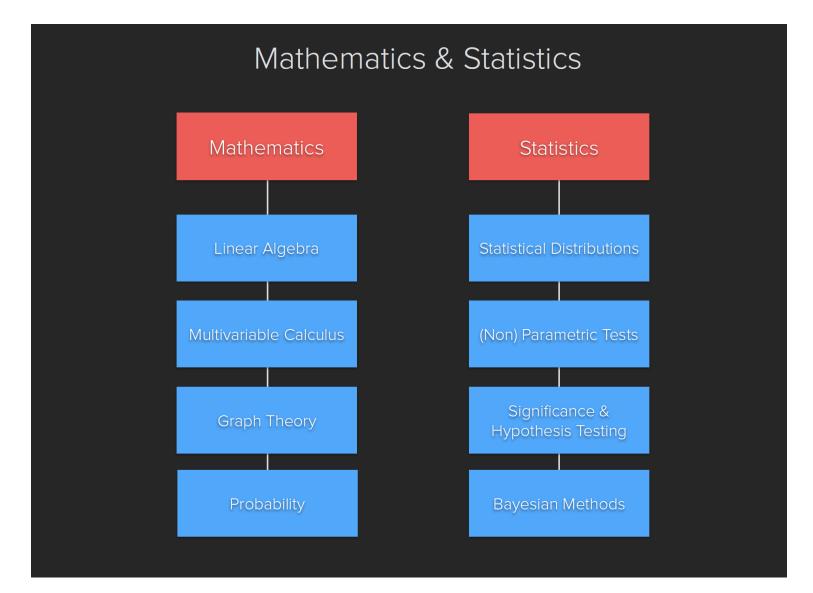
- CS
  - Big data tools
  - Python/R/Scala
- Math/Stats
  - Linear algebra
  - Correlation analysis
  - Hypothesis tests
- Collaboration with domain experts
  - Visualization
  - Prototyping







#### Math/Stats



#### **Online Courses**

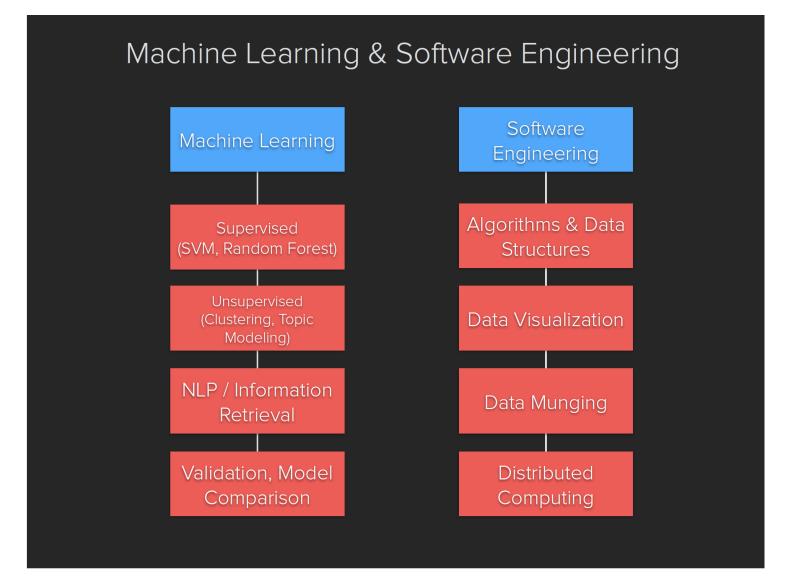
Mathematics & Statistics







#### **Data Analytics**



## **Thank You!**

Good Luck ©



CELEBRATING 30 YEARS Marlan and Rosemary Bourns College of Engineering