WRAP-UP: BIG DATA PROGRAMMING AND MANAGEMENT

Dr. Garrett Dancik

What did we learn?

- Docker
  - Create lightweight, standalone, self-contained executable software packagess
  - Dockerize your Python application:
    - https://runnable.com/docker/python/dockerize-your-python-application
- Linux
  - Free, open source operating systems based on the unix kernel
  - Highly configurable
  - Powerful command line tools
  - Linux *bash* shell can be run on Windows 10: https://itsfoss.com/install-bash-on-windows/

Cloudera and Hadoop

- **Apache Hadoop** (https://hadoop.apache.org/) is "a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models."

  "The name my kid gave a stuffed yellow elephant. Short, relatively easy to spell and pronounce, meaningless, and not used elsewhere. Those are my naming criteria. Kids are good at generating such."
  http://www.balasubramanyamlanka.com/origin-of-the-name-hadoop/

- **Cloudera CDH**, or *Cloudera's Distribution Including Apache Hadoop*, is 100% open source, heavily tested and widely used. (https://www.cloudera.com/)

Real world Cloudera Hadoop examples

- Thompson Reuters analyzes 13 million tweets a day to identify newsworthy events in realtime:
  - https://www.cloudera.com/about/customers/thomson-reuters.html
- HelloFresh uses a Cloudera Data Warehouse and can analyze more than 15 TB of data to make personal recommendations and predictions.
  - http://vision.cloudera.com/how-hellofresh-is-disrupting-the-grocery-industry-using-deep-customer-insights/
- More examples:
  - https://www.cloudera.com/about/customers.html
  - https://aptude.com/blog/entry/5-hadoop-implementation-success-stories/

Cloud platforms that support Hadoop

- Run Hadoop in the Cloud using
  - Microsoft Azure: https://azure.microsoft.com/en-us/services/hdinsight/
  - Amazon EMR: https://aws.amazon.com/emr/features/hadoop/
  - Google Cloud Dataproc: https://cloud.google.com/dataproc/

The Hadoop Ecosystem

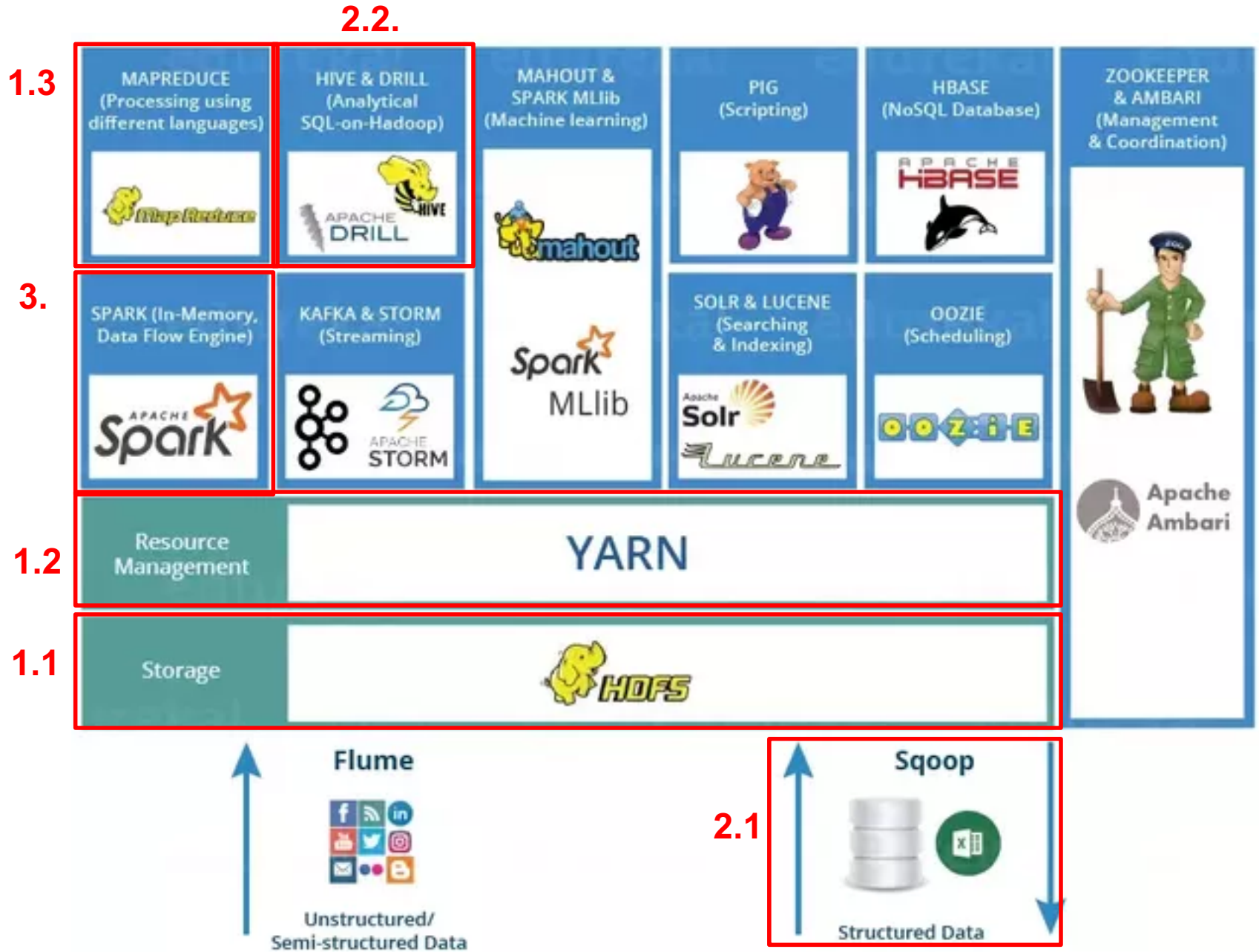

Image posted by Shubham Sinha at quora.com

Additional Tools

- Apache Drill (https://drill.apache.org/)
  - Schema free SQL query engine
  - No table creation; files are queried directly
- Apache Pig (https://pig.apache.org/)
  - High level scripting language and SQL-like scripting language (Pig Latin)
  - Generates one or more MapReduce jobs
- Apache Flume (https://flume.apache.org/)
  - A reliable service for moving streaming data into HDFS

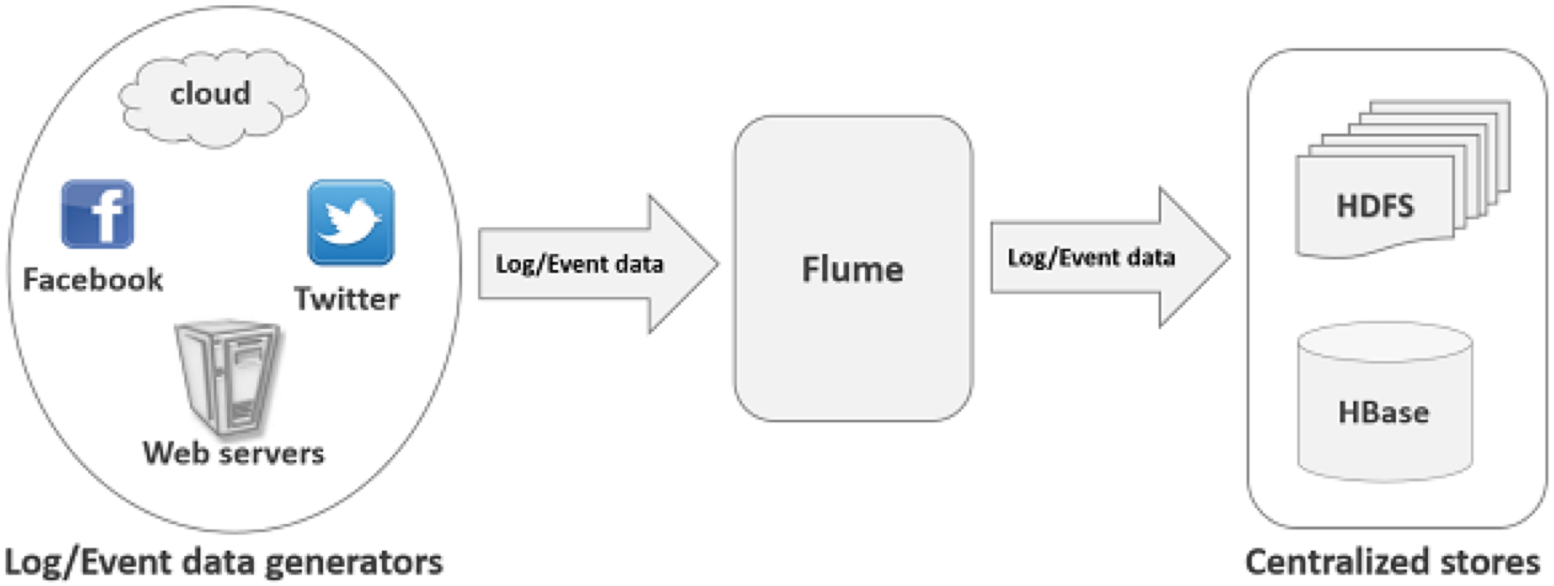

Image: https://www.tutorialspoint.com/apache_flume/images/apache_flume.jpg

Future of Big Data

- Lots of data will continue to be collected (but “Actionable Data”, “Smart Data”, etc may replace “Big Data”)
  - Currently there are more than (https://www.internetlivestats.com/)
    - 6,000 tweets sent out every second
    - 40,000 Google searches every second
    - 4.5 billion FB likes every day
  - Amazon processes ~ 300 transactions per second.
  - The Joint Polar Satellite System (JPSS) takes satellite images covering the entire Earth twice per day.
  - Uses in Healthcare: https://catalyst.nejm.org/big-data-healthcare/
- Technologies are rapidly changing
- 17 Predictions about the future of Big Data that Everyone Should Read: https://www.forbes.com/sites/bernardmarr/2016/03/15/17-predictions-about-the-future-of-big-data-everyone-should-read/