Big Data Engineering

Conclusions and Recap

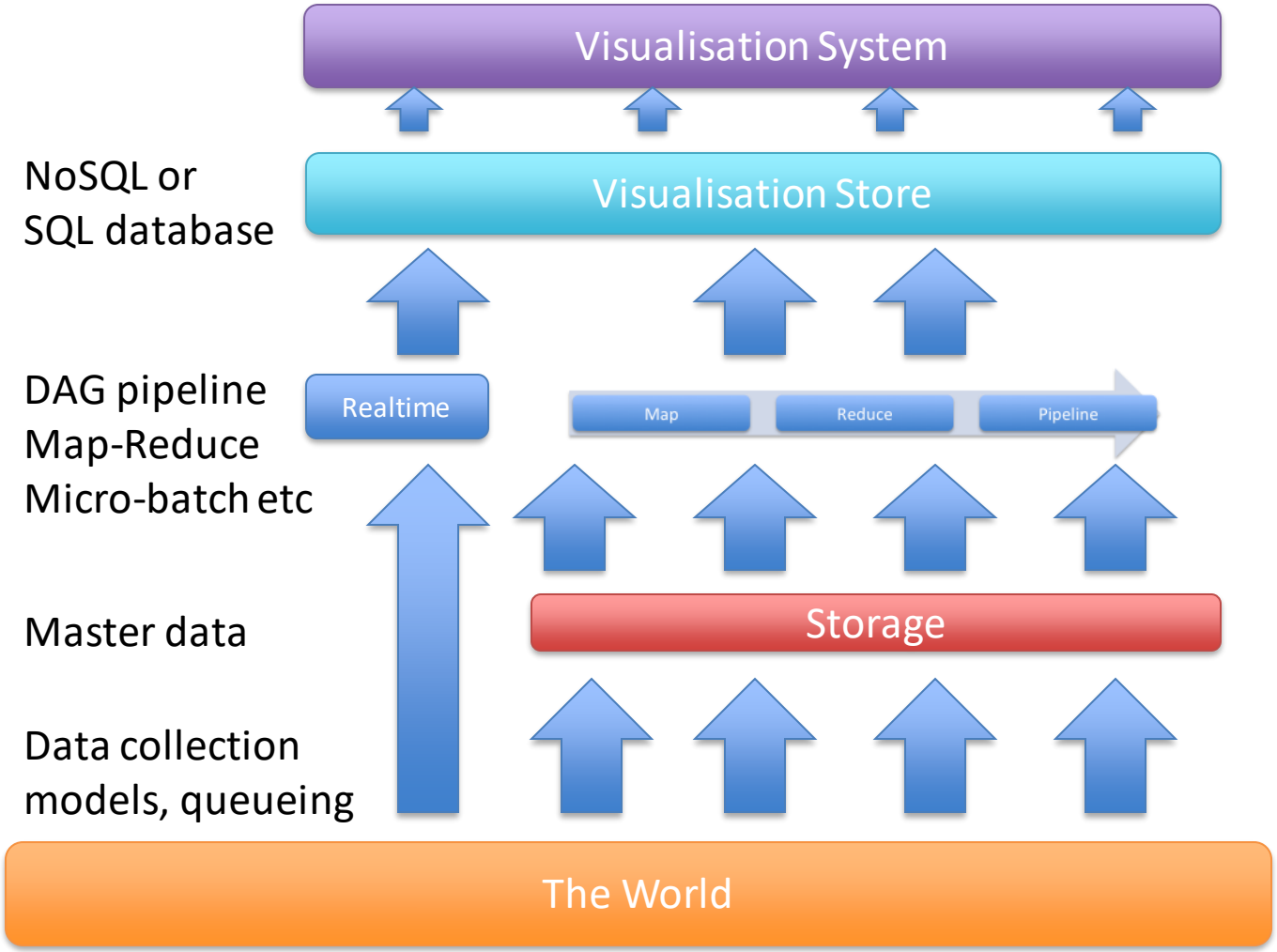
Adam Hill

April 2023

Contents

- Understanding the bigger picture
- What are the different components
- Managing the components
- Message queueing and collection systems
- Map-Reduce and DAG systems
- Realtime Systems
- Fast databases for speed
- Visualisation and Dashboards

The big picture



The big picture

- You have *immutable* master data
- You create a set of processes to:
  - Collect that data
  - Store master data
  - Process data
  - Visualise and present
- Some of those processes act on batch and others on real-time data

How to choose the components?

- Two main approaches:
  - Best of breed
    - Choose the best available component in each space
  - Stack
    - Choose a curated stack that a team or organization is providing/selling/supporting

Approach

- Minimise the pain
  - Choose what you need when you need it
  - Don’t over engineer

Managing to tool/tech ecosystem

- Use Docker!
- Allows you to sandbox different parts of the solution.
- Let's you experiment quickly
- Means that what you build will run anywhere!
- It is a major underpinning technology that together with Kubernetes will allow scaling of solutions.

How do I ingest data?

- File transfer
- Live stream
  - Sockets
  - Syslog
  - Messaging system
- From existing databases

How do I store data?

- HDFS
- zFS / GlusterFS / NFS ...
- NoSQL database
  - Mongo / HBase / Cassandra / Apache Parquet
- CSV

How do I process data?

- Simple Map Reduce
- Hive / Pig
- DAG
- Pipeline
- etc

How do I visualise data

- From a SQL database?
- From a NoSQL database?
- Generate charts in Python Spark?
- Etc?

Collection / Queuing systems

- Two ways of making the choice
  - The protocol
  - The middleware
- Protocols
  - ZeroMQ, MQTT, AMQP, STOMP, Kafka Protocol, Rendevouz, etc
- Middleware
  - Kafka, Apollo, Mosquitto, QPid, WSO2, etc

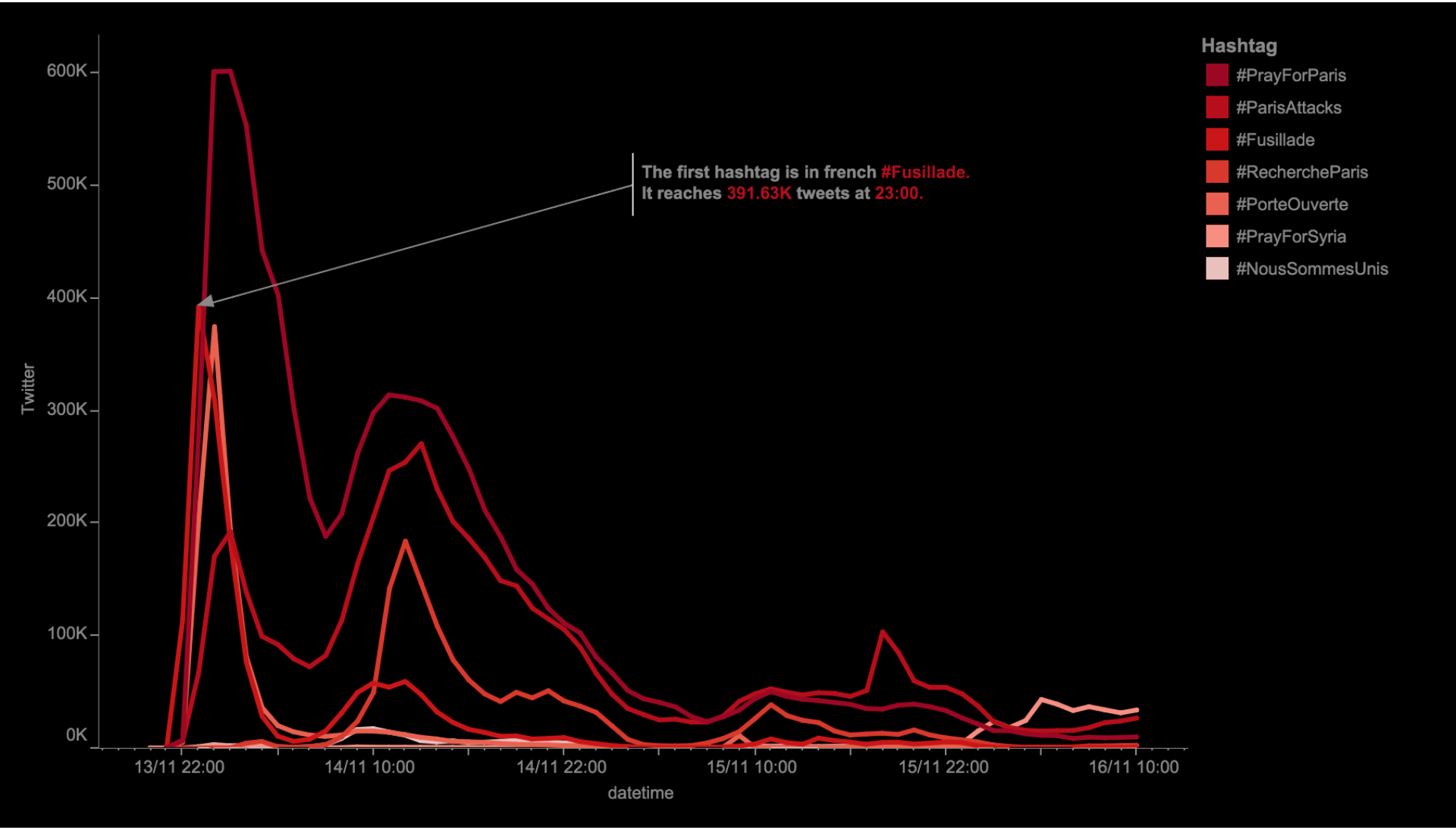
Processing approaches

- Covered in detail already
- Hadoop
- Spark
- etc

Cluster Management

- Spark
- YARN
- Mesos
- Kubernetes
- etc

Visualisation



Visualisation approaches

- Full products
  - Tableau, Qlik, SAS, GoodData
- Web-based systems
  - Tableau Public, Datawrapper, Raw, Plotly
- Developer oriented
  - D3.js, dygraphs, Python charting, Leaflet, Fusion Charts, Google Charts, etc

Fortune top 10 big data companies

fortune.com/2014/06/13/these-big-data-companies-are-ones-to-watch/

- MapR – Apache Hadoop
- MemSQL
- Databricks – Apache Spark
- Platfora – Apache Hadoop
- Splunk
- Teradata – Apache Hadoop
- Palantir – Hadoop, Cassandra, Lucene
- Premise
- Datameer – Apache Hadoop
- Cloudera – Apache Hadoop
- Hortonworks – Apache Hadoop
- MongoDB – MongoDB
- Trifacta – Apache Hadoop

The real answer

You are on the cutting edge

– Expect to have some pain

Questions?

Pandas vs Pyspark

- https://databricks.com/blog/2018/05/03/benchmarking-apache-spark-on-a-single-node-machine.html