

Big Data Engineering

Conclusions and Recap

Julie Weeds

March 2019



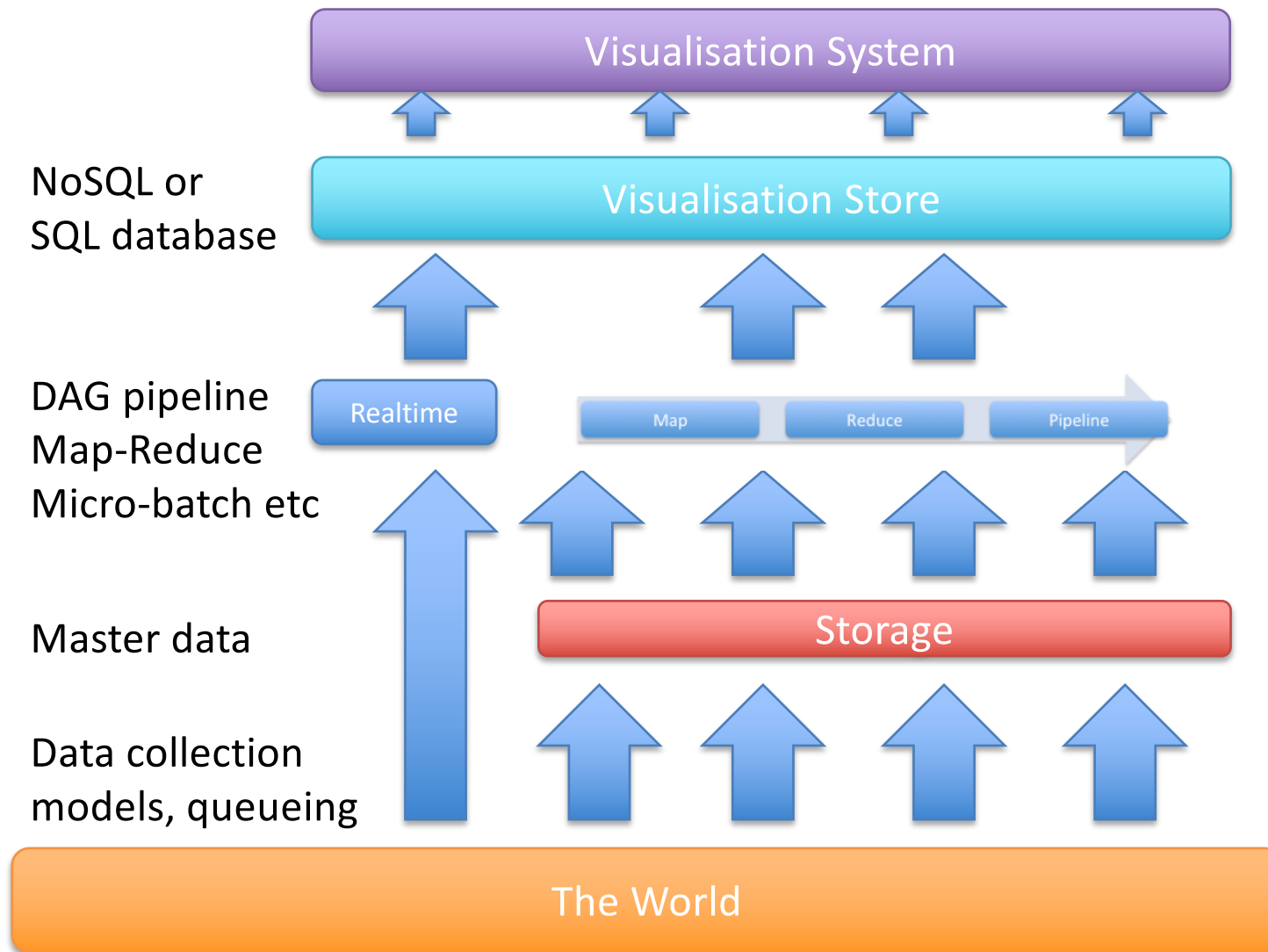
© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>

Contents

- Understanding the bigger picture
- What are the different 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



How do I ingest data?

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



How do I store data?

- HDFS
- NoSQL database only
 - Mongo / HBase / Cassandra
- zFS / GlusterFS / NFS etc
- Apache Parquet, CSV, or speci



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, Rendezvous, etc
- Middleware
 - Kafka, Apollo, Mosquitto, QPid, WSO2, etc



Processing approaches

- Covered in detail already
- Hadoop
- Spark
- Tez
- etc

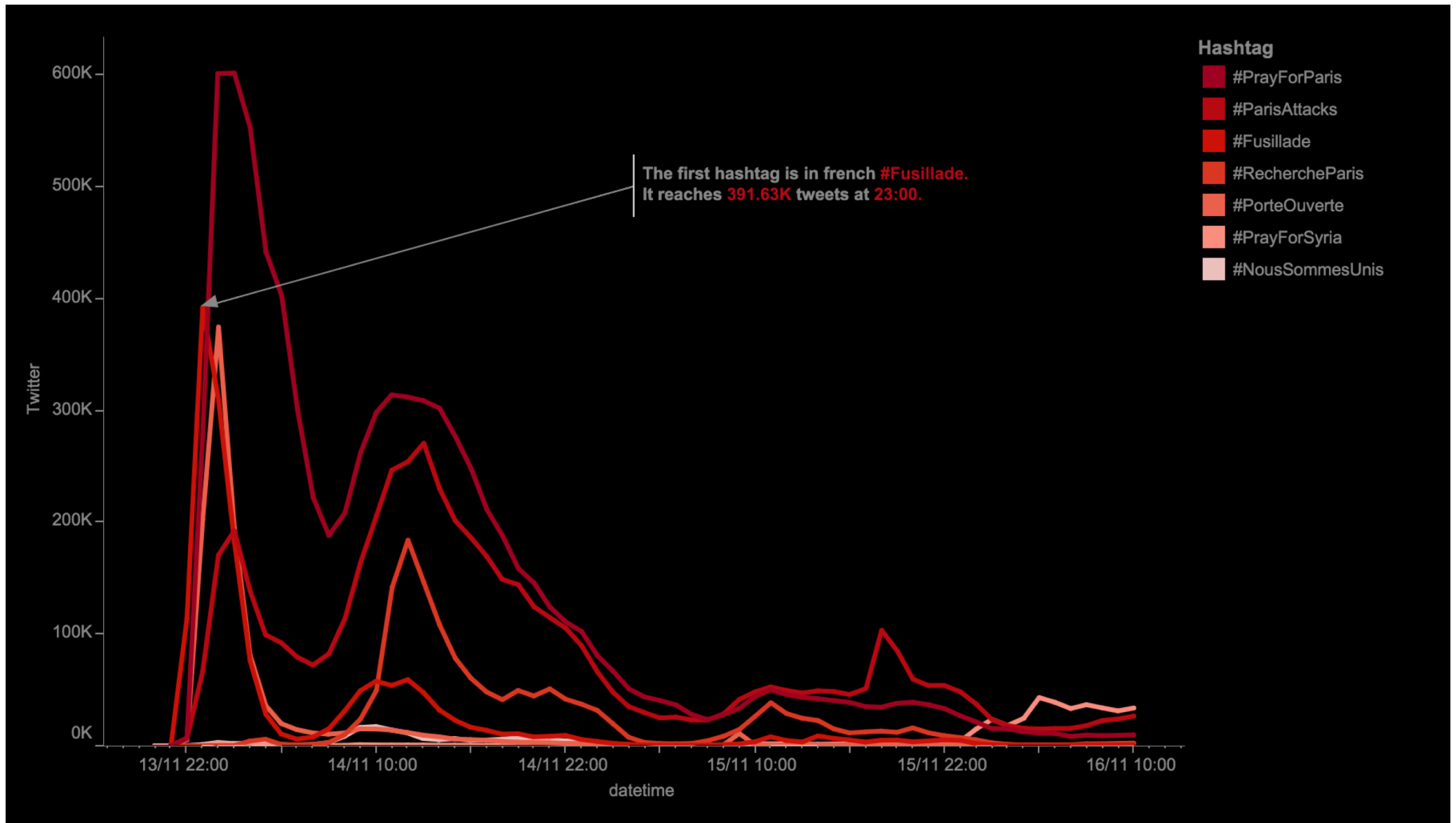


Cluster Management

- 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



Trustradius big data companies to watch 2018

- Business Analytics
 - Alteryx
 - Arcadia Data
 - ClearStory Data
 - Cooladata
- Data management and integration
 - Actian
 - Alation
 - Attunity
 - Iguazio
- Big data platforms
 - **Hortonworks**
 - Micro Focus
 - **Teradata**
- Machine learning
 - DataRobot
 - H2O.ai
 - **Splunk**
 - The Data Incubator
 - Domino



The real answer

You are on the bleeding edge
—Expect to have some pain



Questions?



© Paul Fremantle 2015. This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License
See <http://creativecommons.org/licenses/by-nc-sa/4.0/>