Big Data Engineering

Conclusions and Recap

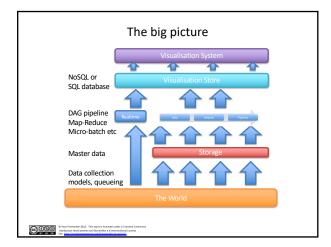
Julie Weeds March 2019

© Paul Fremantie 2015. This work is licensed under a Creative Commo

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

- 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

© 9 00 an Fremantie 2015. This work is Economic under a Greative Co Attribution-NonCommercial-ShareAlike 4.0 international License See point Foresthericommission conficement (hypocopies) of the Commission Comm

Approach

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

© Paul Fremantie 2015. This work is licensed under a Creative Comm
Attribution-NeoCommercial-ShareAlike 4.0 international License

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

© Paul Fremantie 2015. This work is licensed under a Creative Common Attribution-Noof Commercial-ShareAlias 4.0 International License See http://license/secommon.org/licenses/s

How do I process data?

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

© 900 Paul Fremantie 2015. This work is licensed under a Creative Commo Attribution-MonCommercial ShareAlike 4.0 International License

How do I visualise data

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

© Paul Fremantie 2015. This work is licensed under a Creative Common Attribution-NonCommercial-ShareAlike 4.0 International License

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

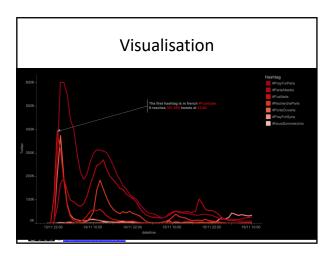
Processing approaches

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

Cluster Management

- Spark
- YARN
- Mesos
- Kubernetes
- etc

© Paul Fremantie 2015. This work is licensed under a Creative Commo



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

Paul Fremandie 2055. This work is licensed under a Creative Comm
 Attribution-NonCommercial-ShareAlike 4.0 international License
 The No. 10 See https://organization.com/picenses/beauto-com/d Ali

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

© Paul Fremantie 2015. This work is licensed under a Creative Commercial-ShareAlike 4.0 International License

Trustradius big data companies to watch 2018

- Business Analytics
 - Alteryx
 - Arcadia Data
 - ClearStory Data
 - Cooladata
- Data management and integration
 - Actian
 - AlationAttunityIguazio

- Big data platforms
 - Hortonworks
 - Micro Focus
- TeradataMachine learning
 - DataRobot
 - H20.aiSplunk
 - The Data Incubator
 - Domino

© Paul Fremantie 2015. This work is licensed under a Creative Commons Attribution-NooCommercial-ShareAlike 4.0 international License

The real answer

You are on the bleeding edge

-Expect to have some pain

© Paul Fremantie 2015. This work is licensed under a Creative Come
Attribution-Neof Commercial-ShareAilise 4.0 international License
See http://exam/vecommercial-coll/fremat/Syneco.ce/4.0.

Questions?	
© 2000 Transactor 2015. This was in Internation and Control Control Community Control Community Control Community Control Community Control Community Control Community Control Control Community Control Control Control Co	