Cloud Computing and Big Data

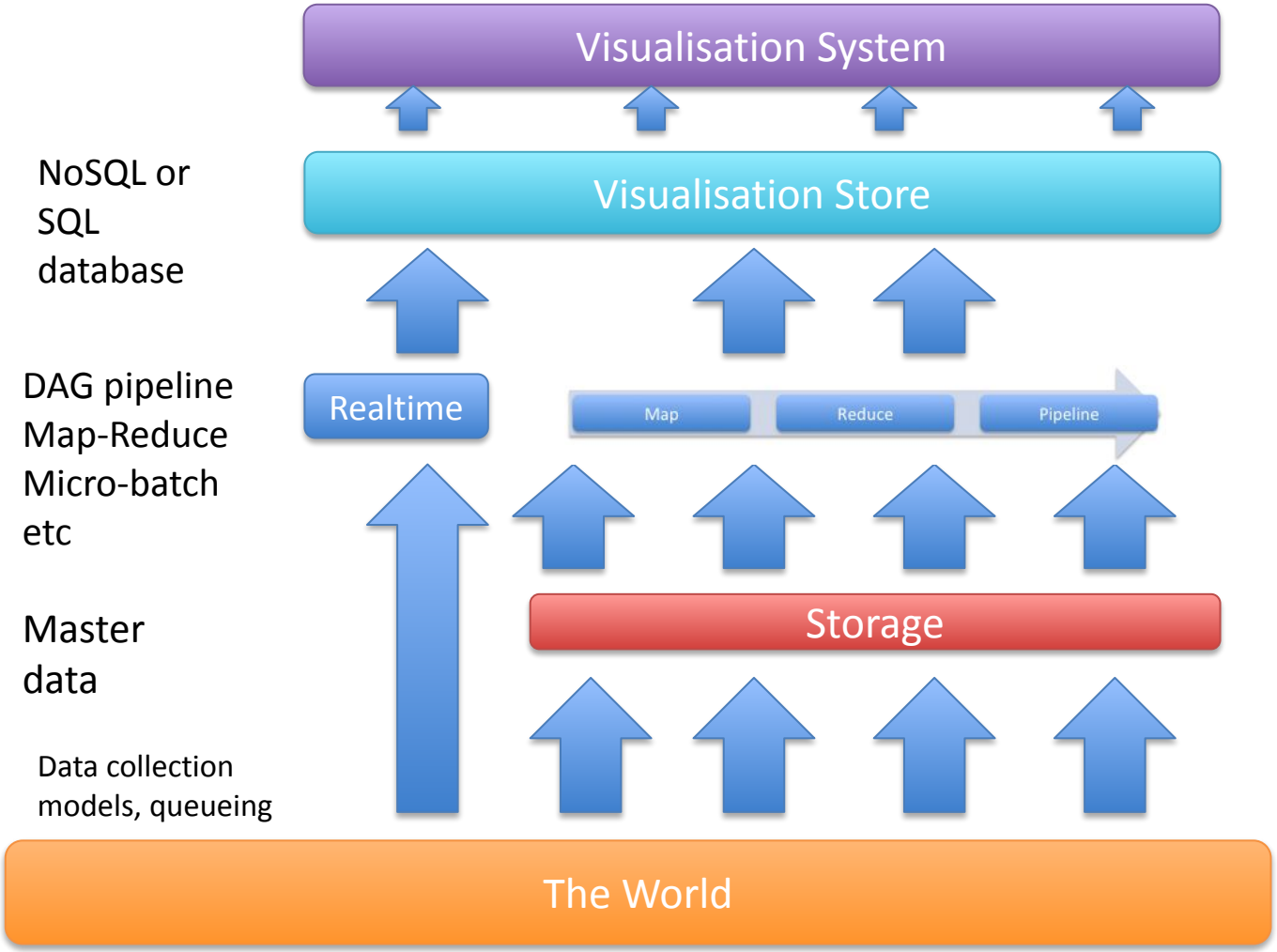
Cloud and Big Data
Pulling it all together

Oxford University
Software Engineering
Programme
July 2020

Contents

- Understanding the bigger picture
- What are the different components
- Message queueing and collection systems
- Map-Reduce and DAG systems
- Realtime Systems
- Theory recap

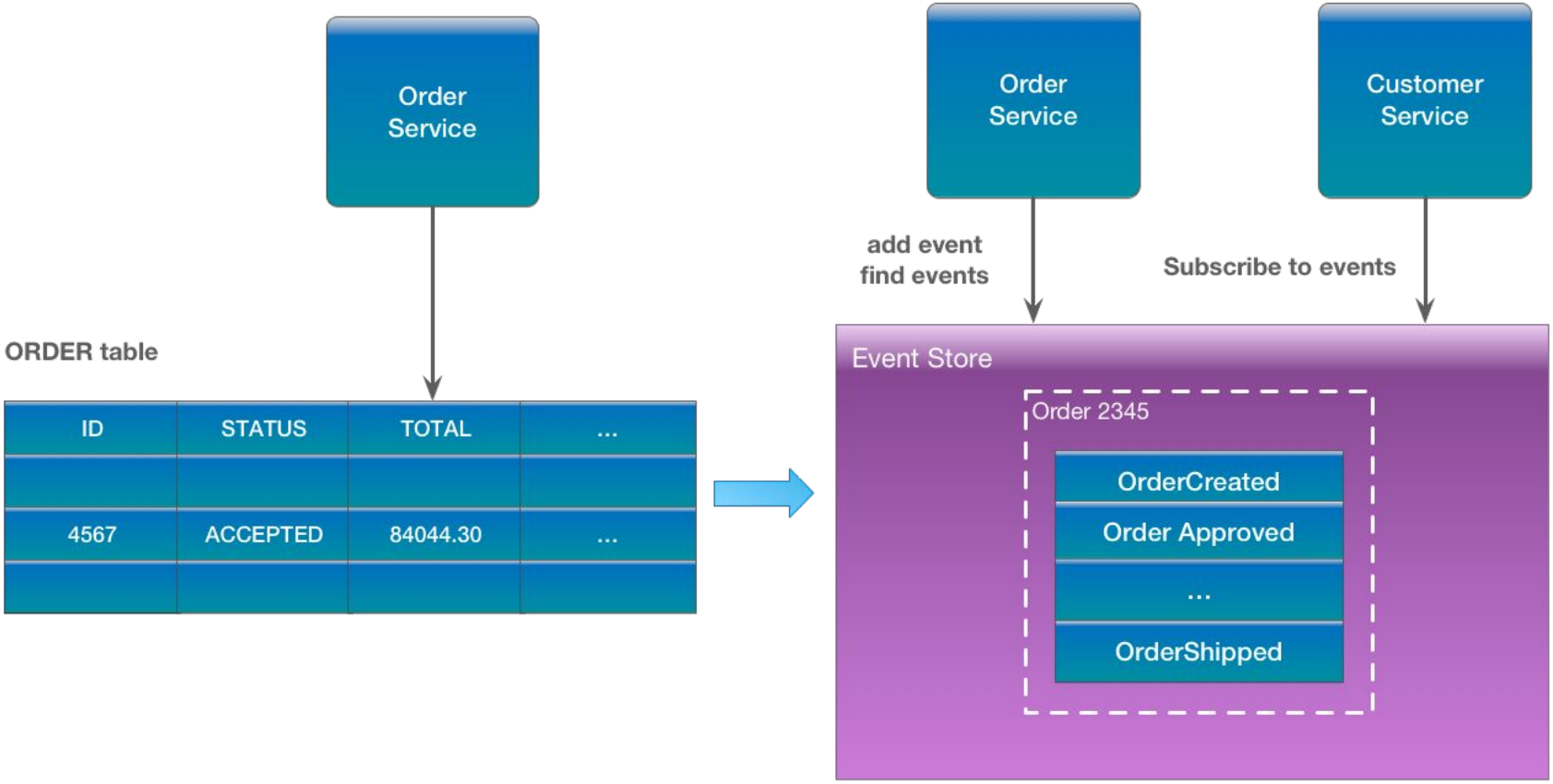
The big picture



The big picture

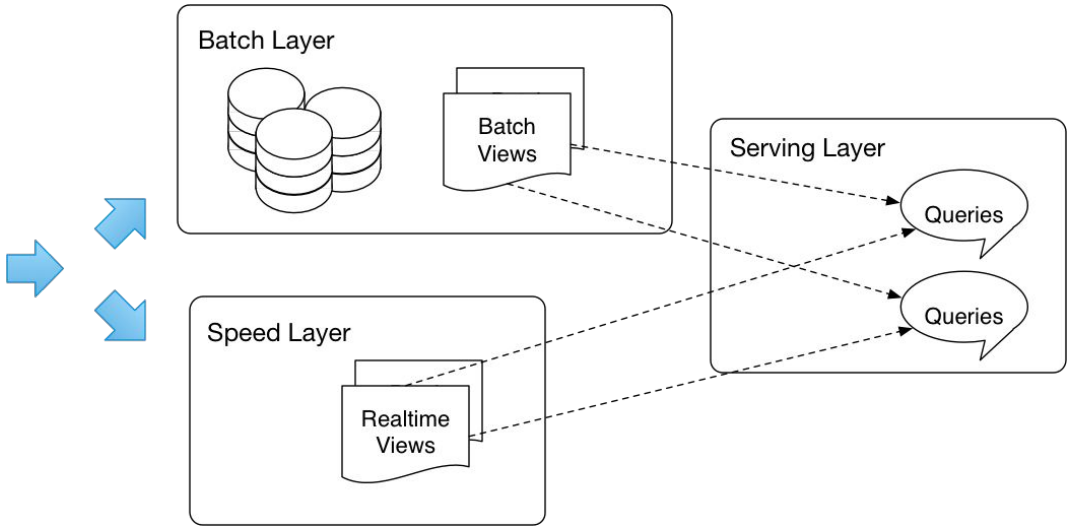
- You have *immutable* master data
- You create a set of processes to:
  - Collect that data
  - Store master data
  - Process data
  - Store aggregates
  - Visualise and present
- Some of those processes act on batch and others on real-time data

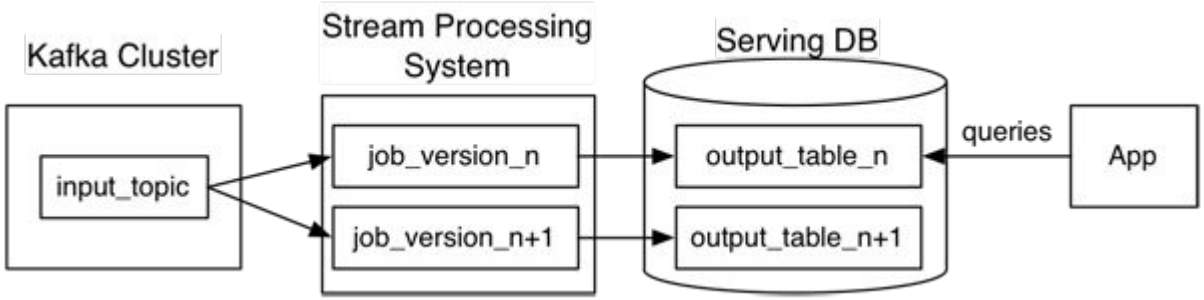
Event Sourcing



https://eventuate.io/whyeventsourcing.html

Lambda vs Kappa





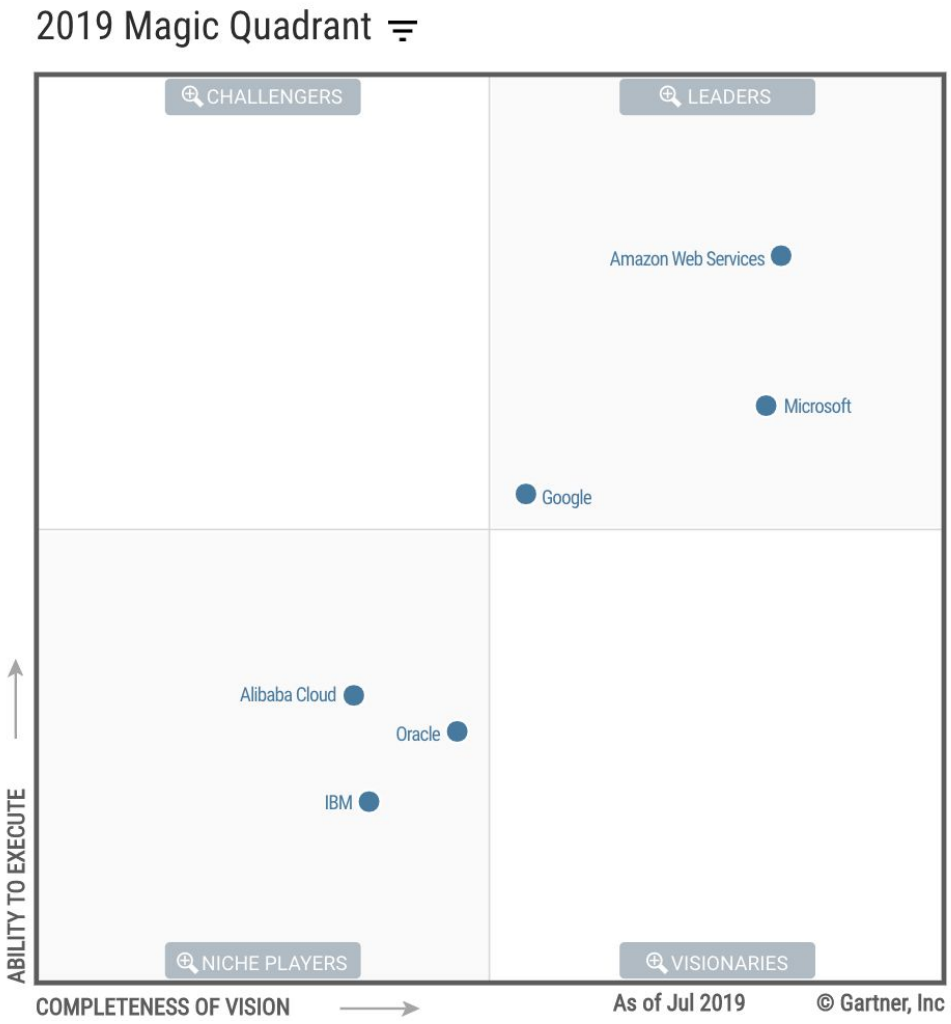
How to choose the components?

- 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

What Cloud Platform?



Market Overview

The market for cloud IaaS is maturing, but revenue is growing unabated. Gartner projects revenue in the cloud IaaS market to increase to $81.5 billion by 2022, up from $41.4 billion in 2019. But most of the enterprise interest and revenue are currently directed toward two providers: AWS and Microsoft. The market views both AWS and Microsoft as being general-purpose providers capable of supporting a broad range of workloads. Google is making steady progress in terms of enterprise adoption, but it remains in a distant third place in terms of overall annual revenue and interest among Gartner's enterprise clients. All other vendors in this market are forced to focus on regional dominance or niche workloads given the momentum of AWS and Microsoft, and the scale at which they operate. Examples of regional and niche-focused vendors are Alibaba and Oracle. Alibaba dominates the market for cloud IaaS in China, and Oracle is, naturally, mostly focused on Oracle workloads as it attempts to scale in the process of rebooting its cloud endeavors. Lastly, IBM remains in a precarious position due to being slow to improve its cloud IaaS offerings, which are ultimately not competitive with the market leaders.

Clip slide

MACRO PERSPECTIVE

MEGACLOUDS ARE FIGHTING TO BE #1 PLUMBING FOR DIGITAL BUSINESS

Besides a few serious regional players like Alibaba, global enterprises have 3 main marketplace bazaars to choose from to power their digital transformation

PLAYER #1 (CATEGORY LEADER): MOMENTUM AND BRAND NAME

$17.1 Billion (2017 Revenue Est.)
40% YoY Growth

PRODUCT STRATEGY
The monocloud that's good enough for most things, not amazing for anything. Heading down proprietary path as most services are integrally tied to their public cloud architecture.

GTM STRATEGY
Aggressive enterprise sales: lock-in, land-and-expand.

BIG EXISTENTIAL QUESTION
Amazon can't allocate 30 top PhDs to solve a single problem. Who will hit Amazon in the achilles heel?

PLAYER #2 (FOR NOW): ENTERPRISE HERITAGE

$6.1 Billion (2017 Revenue Est.)
81% YoY Growth

PRODUCT STRATEGY
Play to internal strengths: Underserved enterprise workloads like legacy Microsoft products, platform and application services for modern enterprise apps.

GTM STRATEGY
Strong enterprise support model.

BIG EXISTENTIAL QUESTION
Will enterprise chops trump Amazon's scale and scope?

PLAYER #3 (KILLER PRODUCTS): BUT WHERE'S THE ENTERPRISE LOVE?

$950 Million (2017 Revenue Est.)
75% YoY Growth

PRODUCT STRATEGY
Google shines strength in machine learning, developer tools, and container orchestration (Kubernetes).

GTM STRATEGY
Historically Google hasn't catered to the enterprise with sales & support. They're apparently trying to change this though.

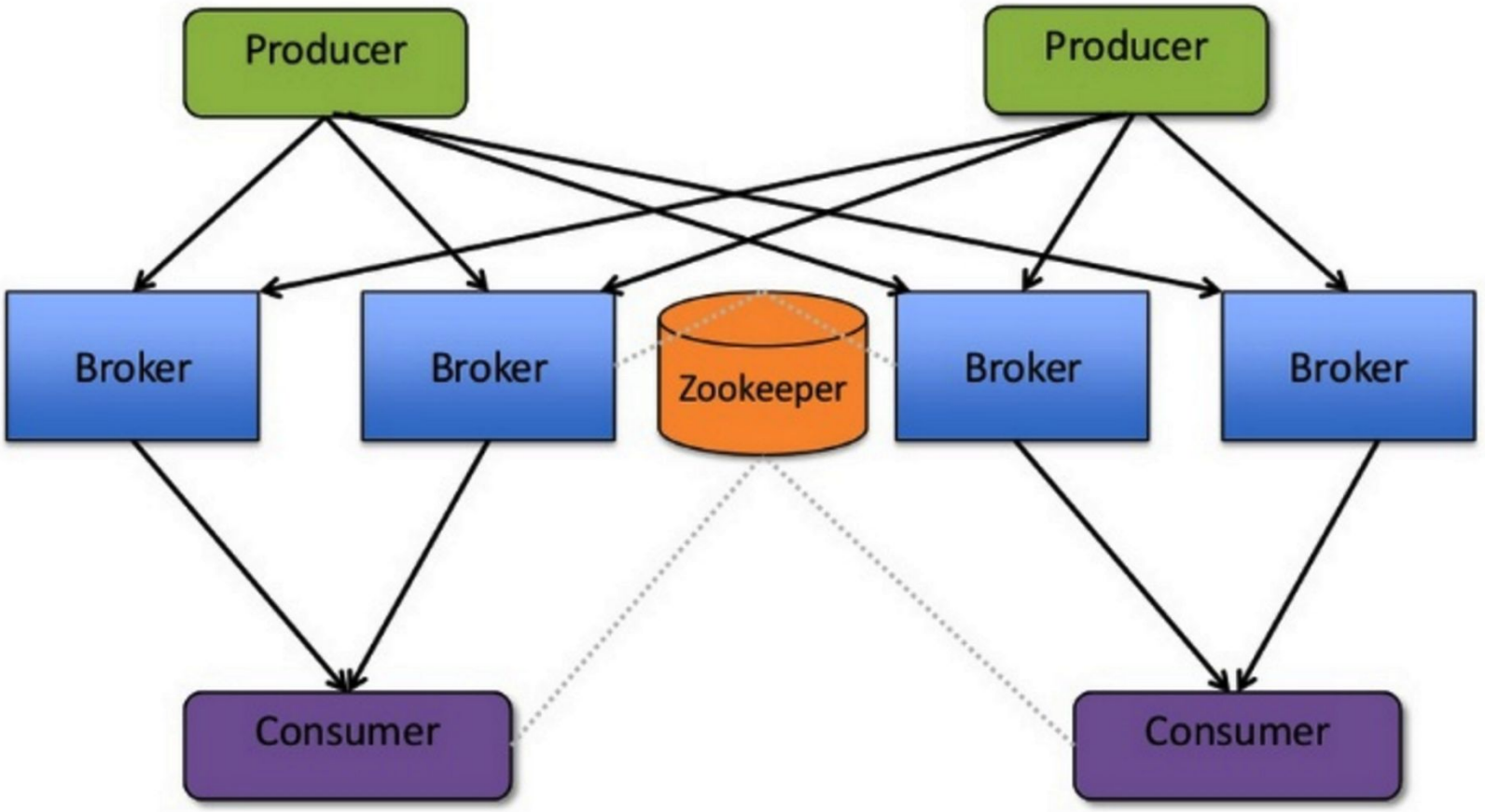
BIG EXISTENTIAL QUESTION
Can Diane Green, Sam Ramji, and the first-class GTM team from Apigee bring Google from enterprise 0 to hero?

Source: Estimates from Bank of America Merrill Lynch's "Server & Enterprise Software: Cloud Wars 9: AI : From faster to smarter powered by ABC." May 8, 2017. Revenue includes PaaS & IaaS.

How do I ingest data?

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

Apache Kafka



Source: http://www.slideshare.net/charmalloc/

How do I store data?

- HDFS
- Cassandra File System
- NoSQL database only
  - Mongo / HBase / Cassandra
- Kafka for Kappa Architecture
- zFS / GlusterFS / NFS etc

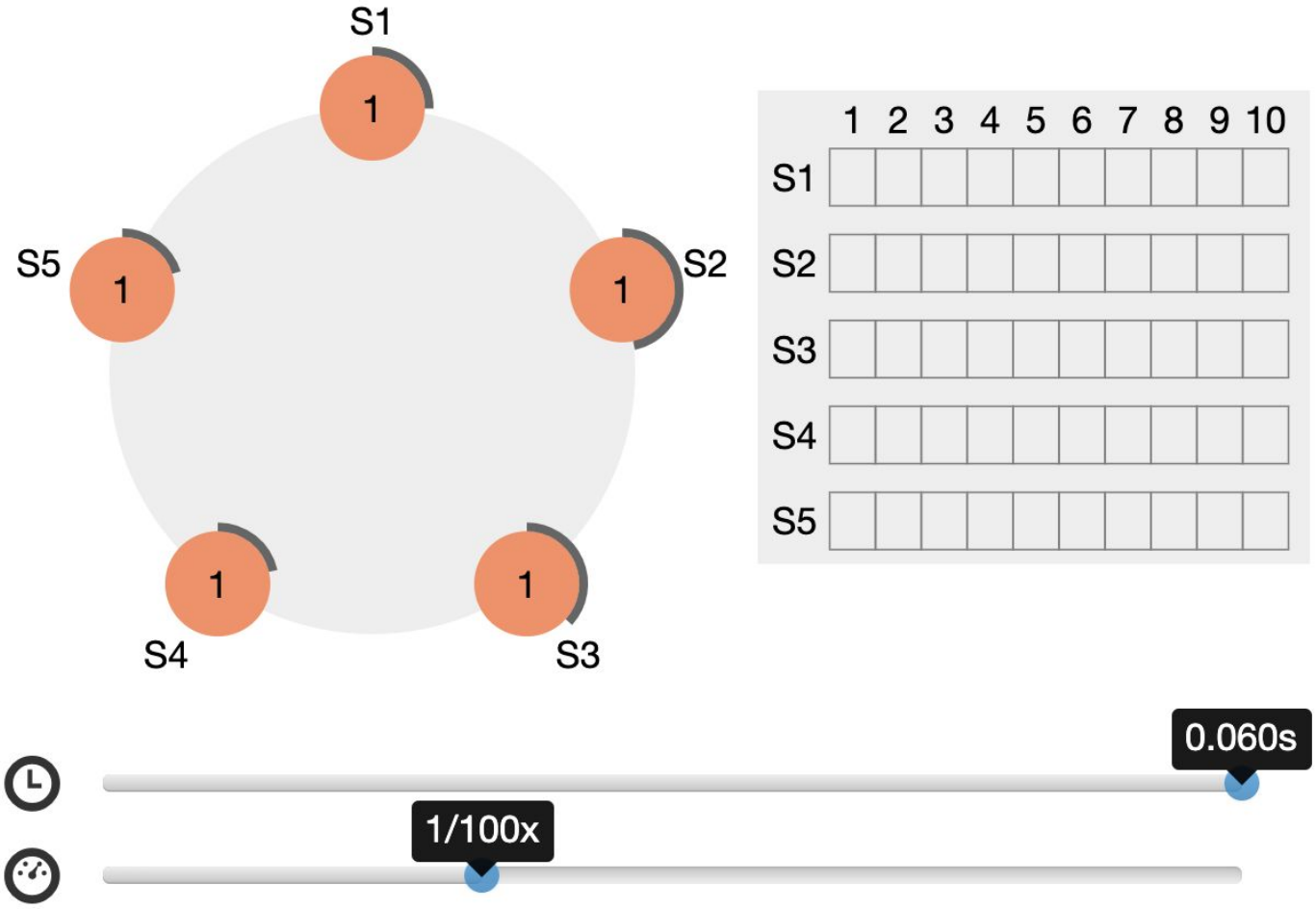
How do I process data?

- Simple Map Reduce (Hadoop)
- DAG
  - e.g Spark, SparkR, SQL
  - Realtime only (Flink, Kafka Streams, Siddhi)
- Pipeline
- etc

Cluster management systems for Big Data

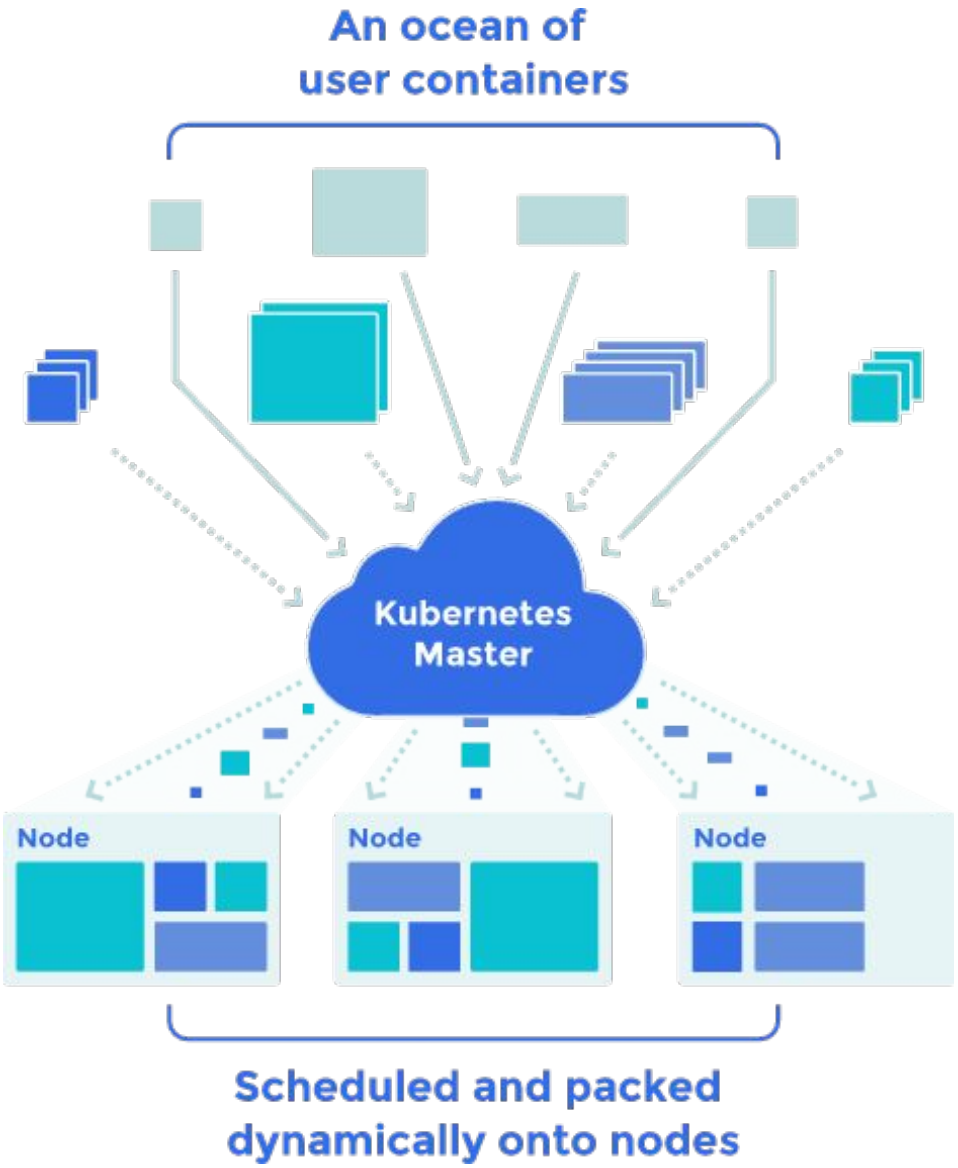
- YARN
- Mesos
- Spark Master
- Kubernetes
  - Is becoming the defacto standard

How do I scale up: Consensus



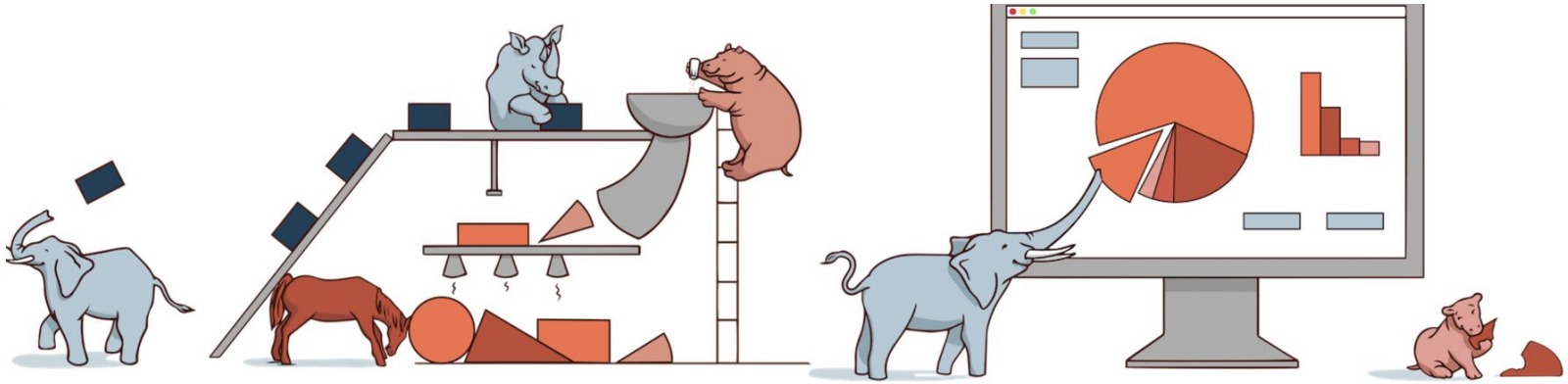
Kubernetes

- An operating system for a datacentre
- “Processes” are high-available scaled containers running in “Pods”



Pachyderm

https://github.com/pachyderm/pachyderm



release v1.5.0 | license apache | godoc reference | go report A+ | Slack Status

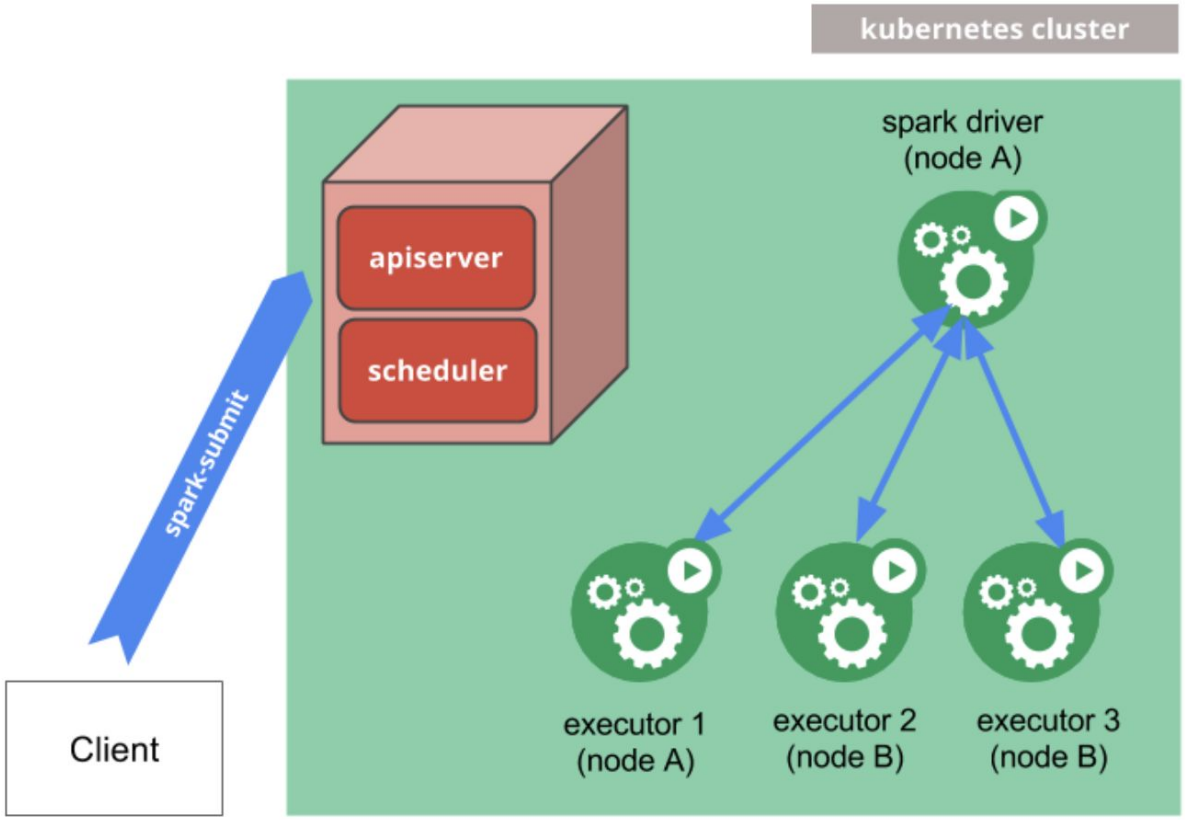
Pachyderm: A Containerized, Version-Controlled Data Lake

Pachyderm is:

- Git for Data Science: Pachyderm offers complete version control for even the largest data sets.
- Containerized: Pachyderm is built on Docker and Kubernetes. Since everything in Pachyderm is a container, data scientists can use any languages or libraries they want (e.g. R, Python, OpenCV, etc).
- Ideal for building machine learning pipelines and ETL workflows: Pachyderm versions and tracks every output directly to the raw input datasets that created it (aka: Provenance).

Spark on Kubernetes

How it works



`spark-submit` can be directly used to submit a Spark application to a Kubernetes cluster. The submission mechanism works as follows:

- Spark creates a Spark driver running within a Kubernetes pod.
- The driver creates executors which are also running within Kubernetes pods and connects to them, and executes application code.
- When the application completes, the executor pods terminate and are cleaned up, but the driver pod persists logs and remains in "completed" state in the Kubernetes API until it's eventually garbage collected or manually cleaned up.

Cassandra on Kubernetes

Oxford University Software Engineering Programme
MSc CLO Module

Exercise 14b

Create a Kubernetes Cluster in DigitalOcean and Deploy Cassandra

Prior Knowledge
Unix Command Line Shell
YAML
Completion of Ex 14a

Learning Objectives
See how Cassandra replicates
Introduction to Kubernetes

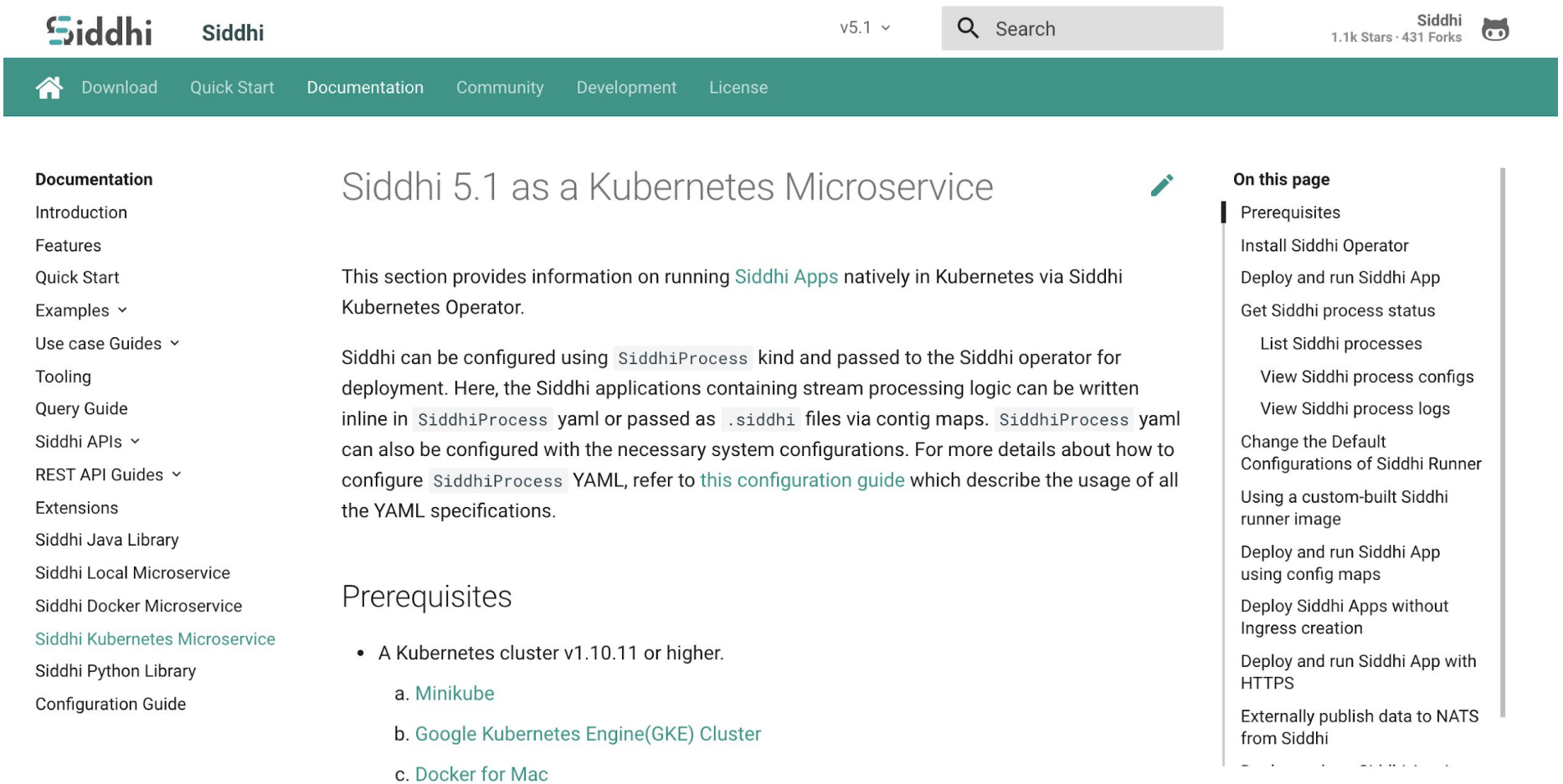
Software Requirements
Browser
kubectl

Realtime

- Apache Storm
  - Highly flexible model
  - Supports pure streaming and micro-batch
  - Lots of plugins
- Apache Spark
  - Micro-batch only
  - Integrates cleanly into Spark (fewer components)
  - Some plugins and more being developed

Siddhi on Kubernetes

https://siddhi.io/en/v5.1/docs/siddhi-as-a-kubernetes-microservice/



A quick recap on theory

- CAP Theorem
  - PACELC as the “solution”
- FLP
  - Raft and Paxos use random timers as the solution
- Scalability at what COST?
- Amdahl’s and Gustafson’s
- Karp-Flatt Metric

Karp-Flatt Metric

e is the Karp-Flatt Metric

ψ is the speedup

p is the number of processors

$$e = \frac{\frac{1}{\psi} - \frac{1}{p}}{1 - \frac{1}{p}}$$

e = 0 is the best

e = 1 indicates no speedup

e > 1 indicates adding processors slows down the system!!!

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

Cloudera and Hortonworks finalize their merger

Frederic Lardinois @fredericl / 6 months ago

Comment

$720m revenue - 2019

Next steps

Education

Students Teachers Schools Events **Get benefits**

Home / Students / GitHub Student Developer Pack

Learn to ship software like a pro.

There's no substitute for hands-on experience. But for most students, real world tools can be cost-prohibitive. That's why we created the GitHub Student Developer Pack with some of our partners and friends: to give students free access to the best developer tools in one place so they can learn by doing.

Free stuff

namecheap

Affordable registration, hosting, and domain management

Benefit 1 year SSL certificate.

Benefit 1 year domain name registration on the .me TLD.

name.com

Domain names, web hosting, and websites. Unicorns and rainbows come standard with our customer support.

Benefit One free domain name and free Advanced Security (SSL, privacy protection, and more).

aws educate

Access to the AWS cloud, free training, and collaboration resources

Benefit Free AWS Educate Starter Account for GitHub Students, worth $100.

Canva

With Canva, anyone can create professional looking graphics and designs. Featuring thousands of templates and an easy to use editor.

Benefit Free 12 month subscription of Canva's Pro tier.

Microsoft Azure

Access to Microsoft Azure cloud services and learning resources – no credit card required

Benefit Free access to 25+ Microsoft Azure cloud services plus $100 in Azure credit.

unity

Accomplish your creative goals using the world's leading real-time development platform, used to create half of the world's games.

Benefit Unity Student Plan free while you are a student.

System availability beyond this week

- Please sign up with **Github Education**
  - See Ex14a
  - Ex14a and b are done using free DigitalOcean credit
  - **You can use free AWS, Azure or DO credit for the assignment**
- What will be running and not!
  - **AWS** will be removed in the next hour
  - **Kafka TFL** until next Friday (Ex13)
  - **Slack** running until Monday - please grab anything you need
- All the materials for the course are always in Github:
  - https://github.com/pzfreo/ox-clo

Thanks!

- I really appreciate everyone's hard work and commitment even when remote!
- Please fill in the feedback forms
- Feel free to add me on LinkedIn https://www.linkedin.com/in/paulfremantle/

But don't message me until you've submitted your assignment!

Questions?