Big Data Overview

Pete Champlin
Population Health Technology
July 14, 2020

KAISER PERMANENTE®

Agenda

- Origins of “Big Data”
- Defining Big Data
- Managing Big Data
- Learning Big Data

 |

KAISER PERMANENTE®

Origins of “Big Data”

A Little History

- Late 1990s/early 2000s
- Google and Doug Cutting - cataloging the World Wide Web to create search engines
- Challenges
  - Massive amounts of data
  - Web pages don’t fit nicely in a table
  - New web pages constantly being created

KAISER PERMANENTE®

Defining Big Data
The Three V's

Volume

- Too much data to fit in one or even several traditional data stores
- Too much data to query quickly or conveniently

Variety

- Data in many different formats
  - Tabular/relational
  - CSV
  - JSON/XML
  - Logs
  - Text (e.g. web pages, emails)
  - Images
  - Audio/Video

Velocity

- Streaming, real-time data
  - Social media/networking (tweets, reviews)
  - Machine generated (logs)
  - Click stream
  - IoT/IoMT (telemetry devices, sensors)
- Real-time data analysis

Managing Big Data

The Three V's

Volume

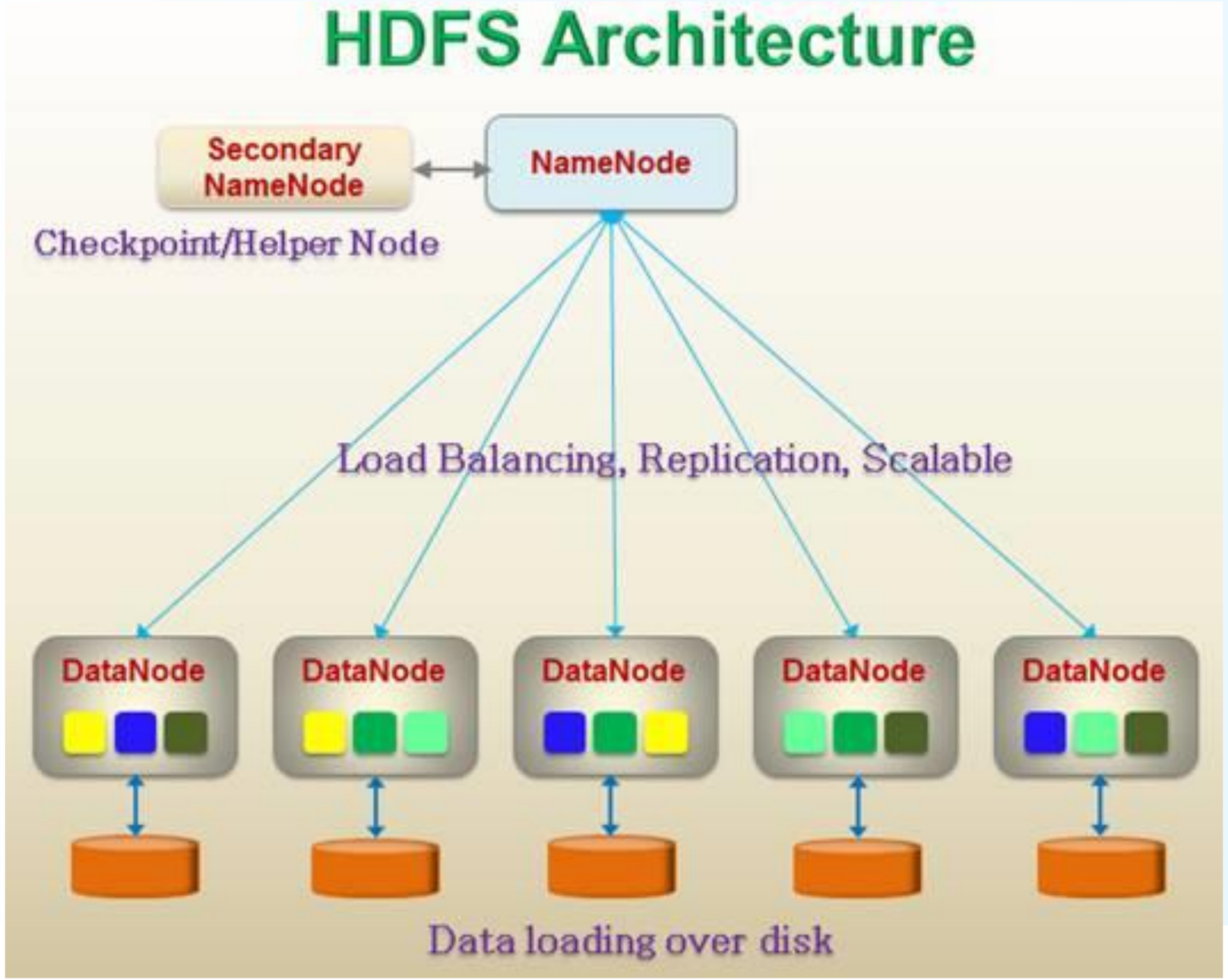
- Horizontal scaling and distributed storage and processing
- Google File System (GFS) & BigTable
- MapReduce processing
- Hadoop Distributed File System (HDFS) & HBase
- Apache Spark
  - In-memory storage
  - Separate processing and storage

Variety

- NoSQL Databases (No SQL or Not Only SQL) – specialize in specific data formats
  - Documents/JSON (MongoDB)
  - Key/Value, column-oriented (Cassandra)
  - Data Warehouse (Snowflake)
- New file formats
  - Parquet
  - ORC
  - Avro

Velocity

- Distributed stream processing systems/engines
  - Apache Storm
  - Apache Kafka (Confluent)
  - Apache Spark (Databricks)



Distributed data and processing tradeoffs - Data volume and velocity issues “solved”, but…

- Hardware failures
- Data movement/shuffling
- Higher latency
- Less (immediate) consistency

 |

KAISER PERMANENTE®

Managing Big Data

Cloud Computing

Storage, processing, and computing services that are remote, distributed, and elastic.

Providers

- Amazon Web Services (AWS)
- Microsoft Azure
- Google Cloud Platform (GCP)
- The rest…Alibaba, IBM, Oracle

Models

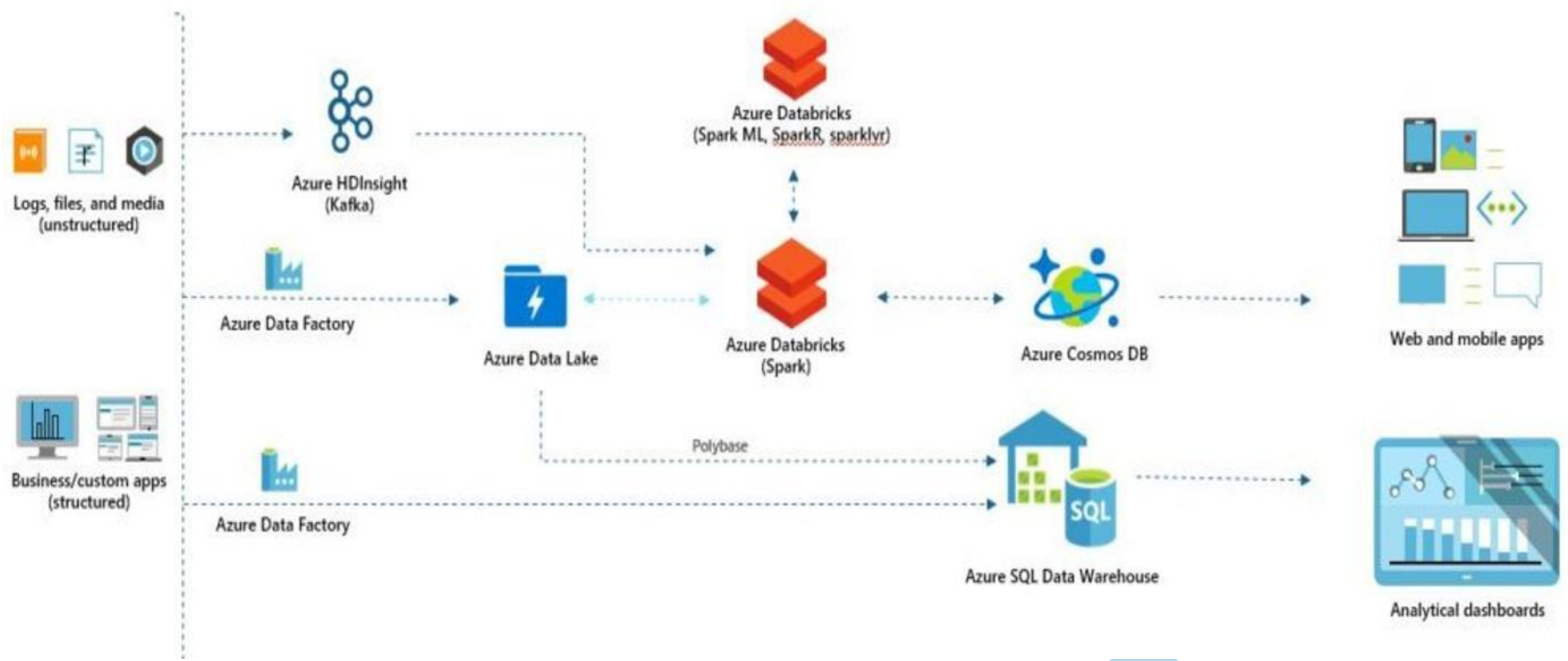
- IaaS (Infrastructure as a Service)
- PaaS (Platform as a Service)
- SaaS (Software as a Service)
- Hybrid (cloud and on-prem)
- Multi-cloud

Benefits

- Scalability
- Managed hardware maintenance
- Evergreen environments
- Data redundancy
- Pay per usage
- Flexibility
- Etc…

Managing Big Data

Architecture



Who Works With Big Data?

Software Engineer

Data Engineer/Architect

Cloud System Analyst/Engineer/Architect

DevOps Engineer

BI Developer/Analyst

Data Analyst

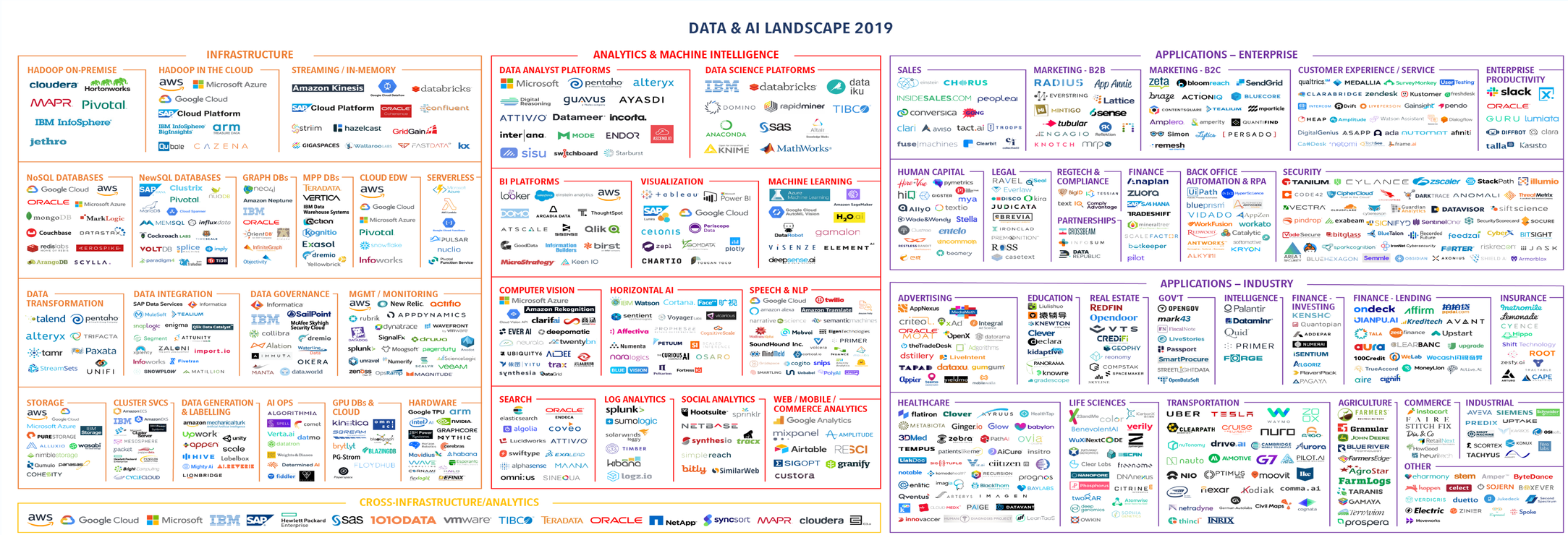
Machine Learning Engineer

Data Scientist

AI *Engineer

KAISER PERMANENTE®

Learning Big Data



July 16, 2019 - FINAL 2019 VERSION

© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap) mattturck.com/data2019

FIRSTMARK EARLY STAGE VENTURE CAPITAL

KAISER PERMANENTE®

Learning Big Data

- **SQL**
- **Python (NumPy, Pandas, Scikit-learn, Anaconda)**
- **Apache Spark/Databricks**
  - Spark is open source, but managed by Databricks (on Azure or AWS)
  - Community Edition (free) - community.cloud.databricks.com

    - Cloud services
    - Machine Learning
    - Notebooks (Jupyter, etc.)
    - NoSQL databases (MongoDB, Cassandra, Snowflake, etc.)

      - R, Scala, Bash, Apache Kafka, Docker/Kubernetes, GitHub/GitLab, CI/CD

KAISER PERMANENTE®

Questions?

KAISER PERMANENTE®