BIG DATA ARCHITECTURES

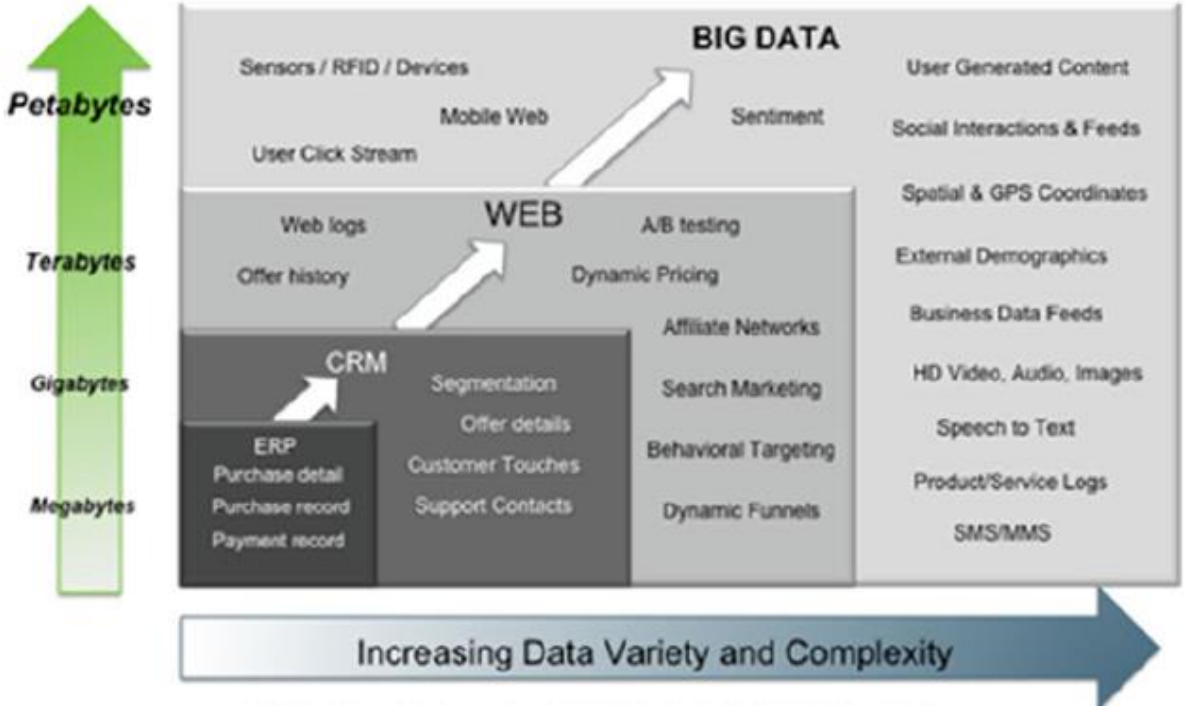
Usama Kaleem, Hyejo Hwang
AIM3 - Scalable Data Science
DIMA / TU-Berlin
09.06.2017

Index

- Introduction
- Definition
- Properties of Big Data Architecture
- Architecture types
  - Lambda Architecture
  - Kappa Architecture
  - Cloud-based Architecture
- Major components
- Challenges & Issues

Introduction

- Big data volume started growing with user interactions



- 10% : structured
- 90% : unstructured

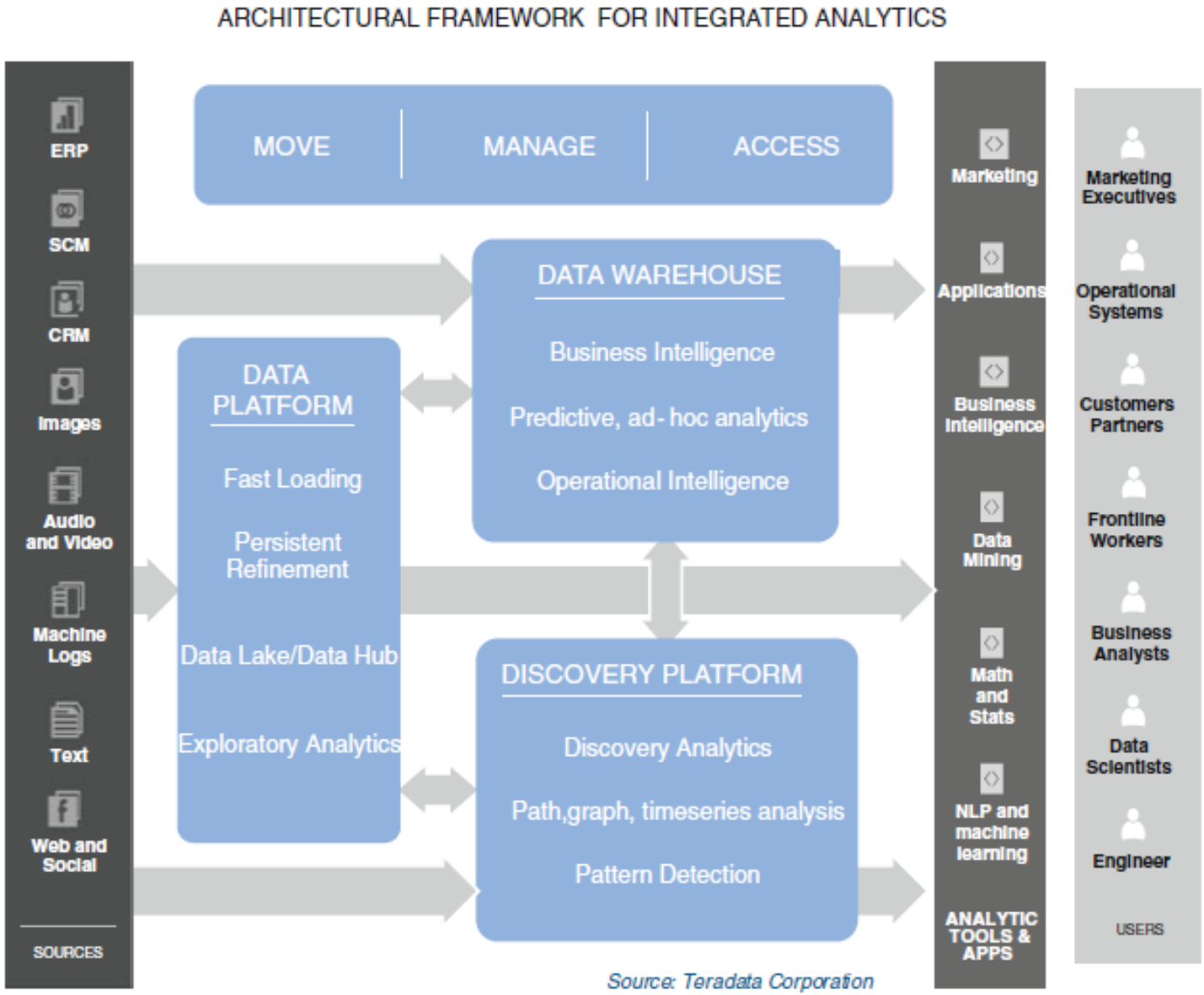
  like emails, videos,

  website clicks..

Source: H. Mohanty et al. (eds.), Big Data, Studies in Big Data 11

Three functional aspects to big data

- Data lake
- Data product
- Data R&D



Source: H. Mohanty et al. (eds.), Big Data, Studies in Big Data 11

Definition

“Big data architecture is an architecture that provides the framework for reasoning with all forms of data.”

Properties of Big Data Architecture

- Scalability
- Resiliency
- Fault-tolerance
- Energy efficiency
- Security
- Operational costs

Architectures Types

- Lambda, Kappa architectures
- Cloud Based architectures
- Streaming architectures
- Parallel and Distributed System architectures
- Novel Hardware and Software Architectures

Lambda Architecture

- Presented by Nathan marz
- Handles massive quantities of data
- Supports real time processing

Batch layer

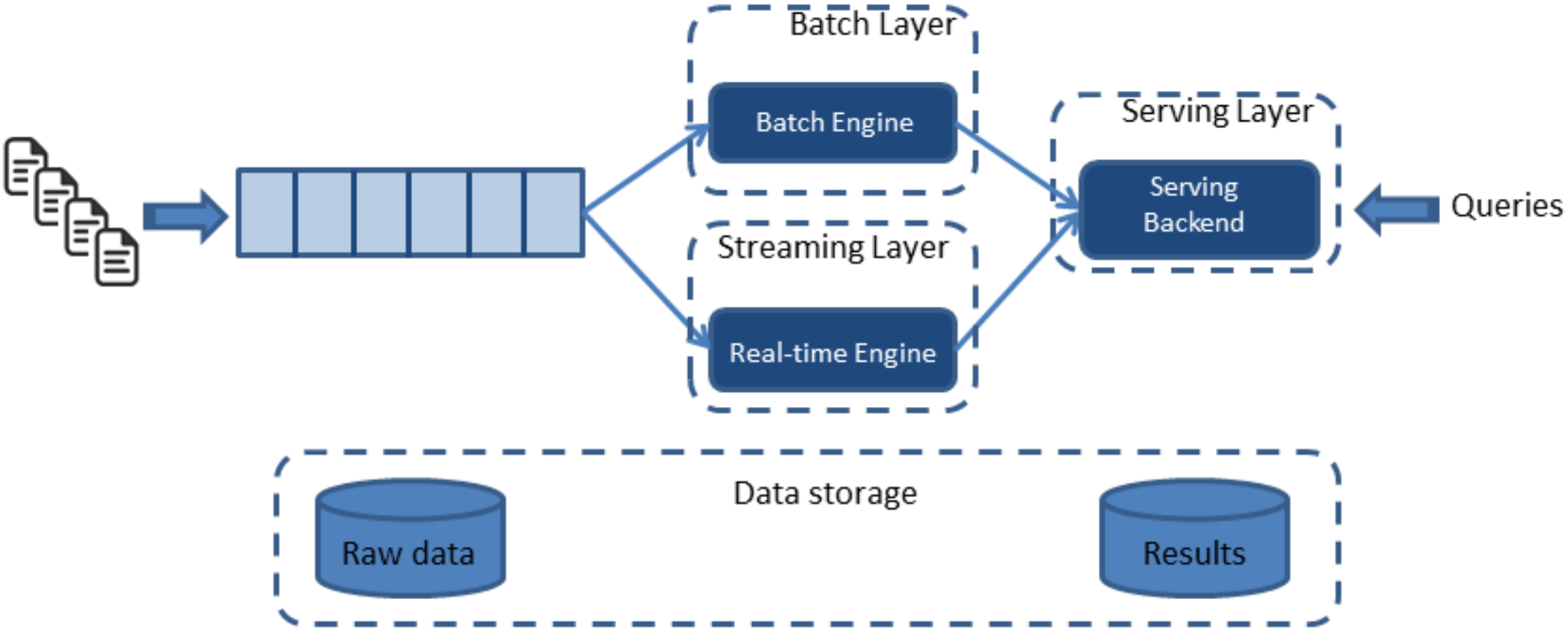
- manages the master dataset, an immutable, append-only set of raw data
- precomputes arbitrary query functions, called batch views
- fares well in resilience aspect

Streaming layer

- processes data streams in real time

Serving layer

- aggregates and merges computation results from both layers



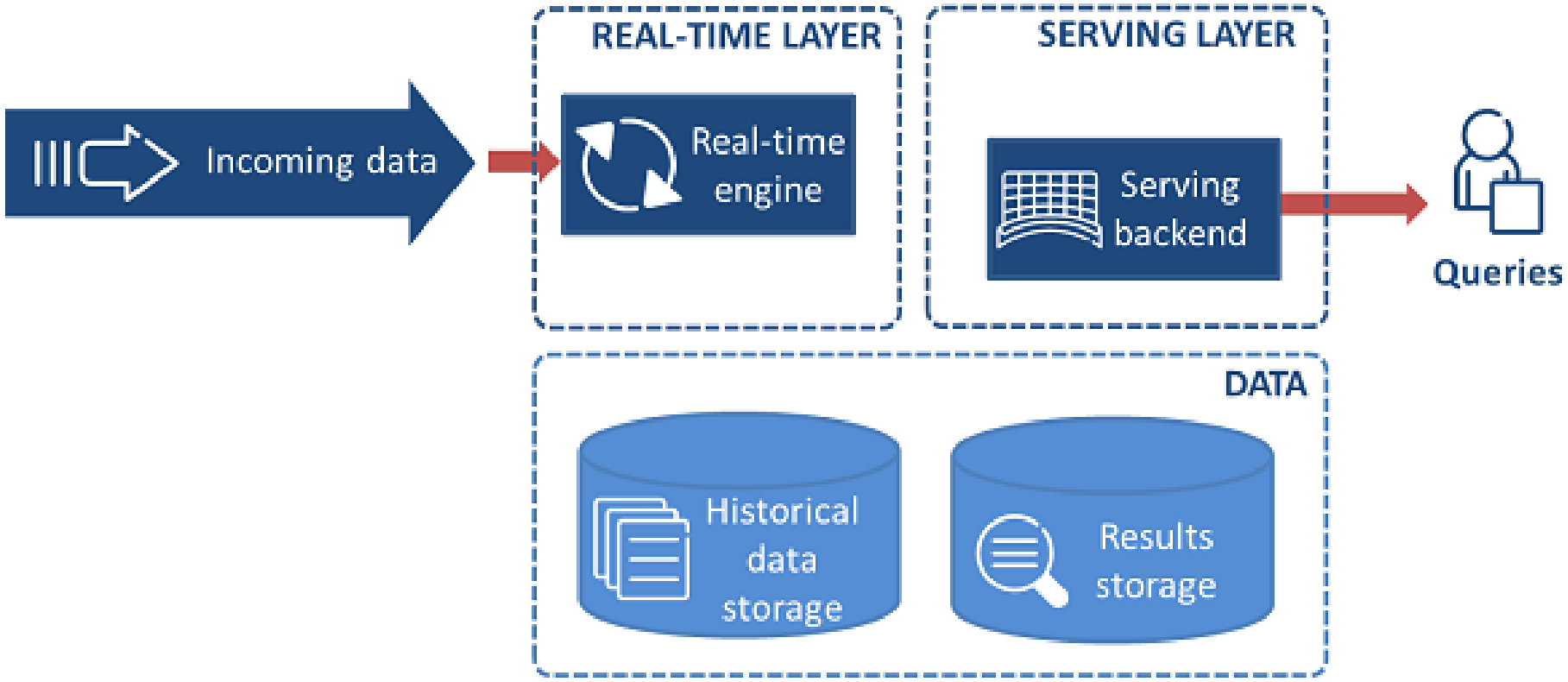
Source : www.oreilly.com/ideas/applying-the-kappa-architecture-in-the-telco-industry

Properties of Lambda Architecture

- If the real-time layer fails, no data will be lost
- Keeps the raw information forever
- Balances latency, throughput, and fault-tolerance
- High complexity

Kappa Architecture

- Suggested by Jay Kreps
- Simplifies the Lambda architecture
- Avoids maintaining two separate code bases for the batch and streaming layers
- Uses a single stream processing engine to handle both real-time data processing and continuous data reprocessing



Source: https://www.ericsson.com/research-blog/data-knowledge/data-processing-architectures-lambda-and-kappa

Streaming layer

- Processes data streams in real time
- Data reprocessing

Serving layer

- Used to Query the results

Properties of Kappa Architecture

- Everything can be treated as a stream
- Immutable data sources
- Replay functionality

Cloud Based Architectures

- Cloud market for big data growing rapidly.
- Promises the reduction of CapEX and OpEx.
- The variety of possible configurations and pricing schemes makes it difficult for consumers to estimate overall costs of cloud-based big data applications.
- Consumers can be cloud application providers themselves e.g. Netflix and Spotify.

Generic Reference Architecture

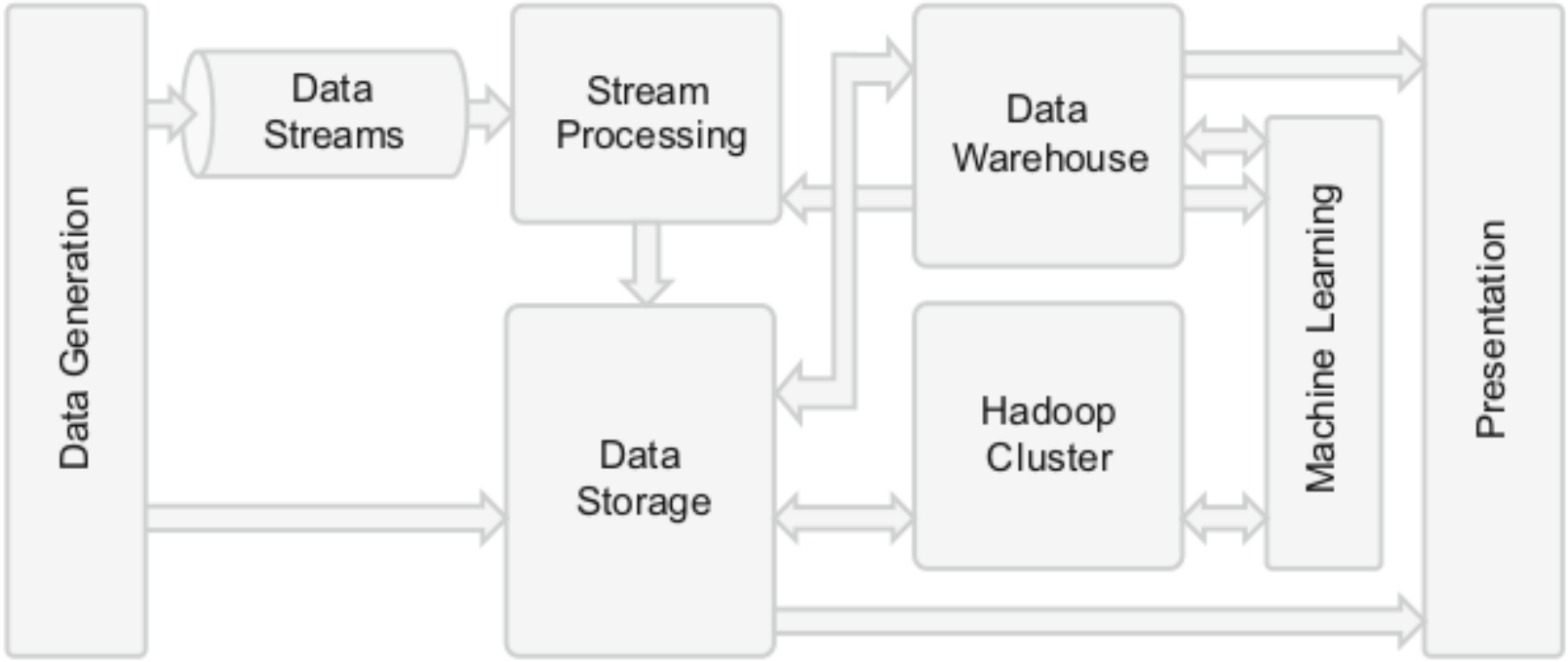


Fig. 2 Generic reference architecture

Managing Cloud-Based Big Data Platforms

Gartner's Leading Three Cloud Providers

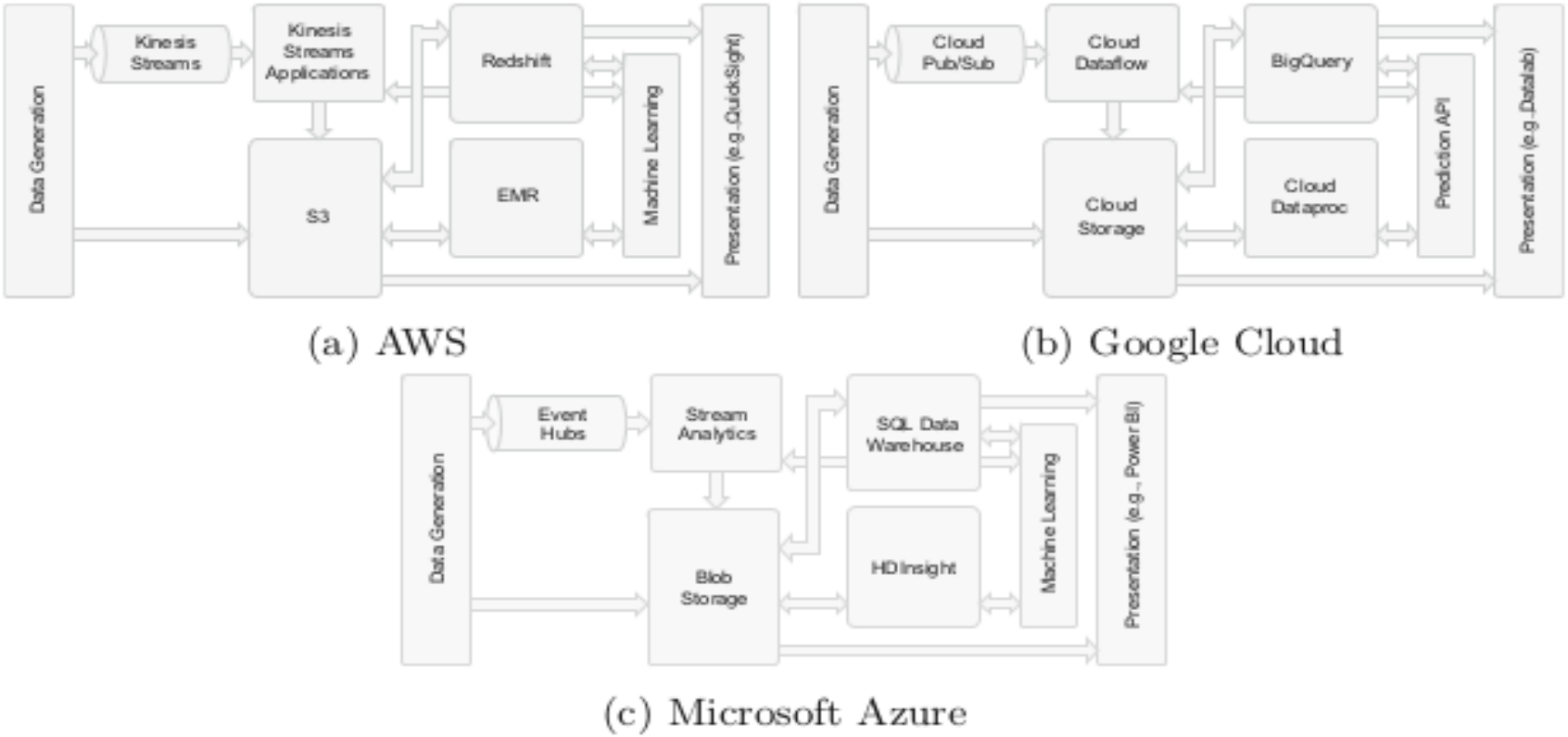
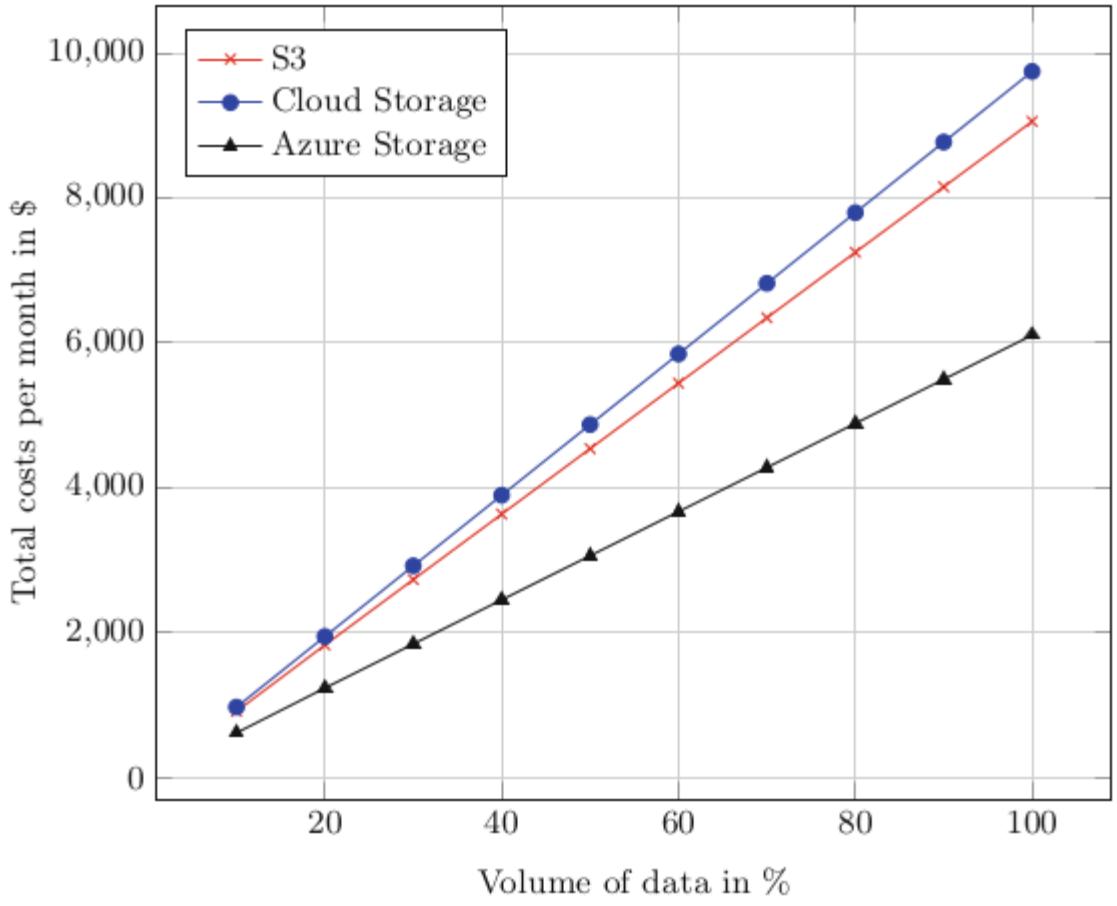


Fig. 3 Reference architectures of leading cloud service providers

Managing Cloud-Based Big Data Platforms

Data Volume Cost



Managing Cloud-Based Big Data Platforms

State-of-the-art

- Systems that support large volumes of both structured and unstructured data is rising continually
- Data is being deployed on cloud services
- Demand for analytical tools is growing that seamlessly connect to and combine a wide variety of cloud-hosted data sources.

Major Components

- Data Generation
- Data Ingestion
- Data Storage
- Data Analytics
- Data Visualization

Managing Cloud-Based Big Data Platforms

Data Generation

- The number of data producers and the amount of data being produced is continuously increasing.
- Data from internal systems (e.g., production data, inventory data, sales data, e-commerce platform data, etc.) and data from external third-party systems (e.g., social network data, government data, weather data, finance data, search trends, etc.) being offered through the Internet.
- The emergence of the internet of things (IoT), enabling physical objects to sense.

Data Ingestion

- Essential to consider velocity requirements I.e how fast the data is fed and processing latency.
- System consists of both streaming and batch processing.

- Stream processing component contains all logic to immediately utilize streaming data, for instance, by generating alerts or making recommendations based on machine learning tools
- May involve ETL.

Data Storage.

- Data should be persisted as files or data records.
- Storage options include: a traditional data warehouse; a data lake; a distributed/cloud-based storage system; and your company server or a computer hard disk.
- cloud-based storage is a brilliant option for most businesses.
- For data analytics in a data warehouse, it is necessary to load the data into database tables of the data warehouse using ETL.

Data Analytics

- Turning data into insights
- Process and analyze the data you have stored.
- Three basic steps in this process

1. Preparing the data
2. Building the analytical model
3. Drawing conclusion from the insights gained

- Softwares by vendors IBM, Google, Oracle for this purposes e.g. BigQuery, Cloudera, Microsoft HDInsight and Amazon Web Services.

Data Visualization

- Once the data has been analyzed and stored, resulting insights and results need to be transformed into rich visualizations.
- Data for stakeholders in form of reports, charts, figures and key recommendations.
- Tools: Management dashboards, word clouds

Architectural Challenges & Issues

- Proliferation of Tools
- Privacy
- Nonproprietary Data
- Massive Storage Requirements
- Cost Reduction
- Automating Test Generation
- Supporting Heterogeneous Hardware Platforms
- Generating Realistic Mixed Workloads
- Assimilating inconsistent component
- Systems must be continually adapted

Conclusion

- The future of data processing is unbounded data.
- Requires a fundamental shift of approach.
- It's difficult to take a decision and come up with a system which acts ideal for all the axis I.e. latency, correctness, cost.
- For a real-world system, evolution is unavoidable and no system can survive such rapid environmental change without evolving correspondingly.


ANY
QUESTIONS
?

References

- [1] Big Data Architecture Goals and Challenges, Cipson Jose Chiriyankandath
- [2] Managing Cloud-Based Big Data Platforms: A Reference Architecture and Cost Perspective, Leonard Heilig and Stefan Voß
- [3] Big Data Provenance: State-Of-The-Art Analysis and Emerging Research Challenges, Alfredo Cuzzocrea DIA Department University of Trieste and ICAR-CNR Italy
- [4] Benchmarking Big Data Systems: State-of-the-Art and Future Directions Rui Han, Zhen Jia, Wanling Gao, Xinhui Tian, and Lei Wang
- [5] Architectural Challenges of Ultra Large Scale Systems, Mehdi Mirakhorli, Amir Azim Sharifloo, Fedeidoon Shams
- [6] Big Data, H. Mohanty, Big Data 11
- [7] Bic Data Architecture Evolution: 2014 and Beyond, Atif Mohammad, Hamid Mcheick, Emanuel Gran

- [8] https://casmodeling.springeropen.com/articles/10.1186/s40294-015-0012-5
- [9] https://www.tableau.com/resource/top-10-big-data-trends-2017
- [10] Big Data Processing Systems: State-of-the-art and Open Challenges, Fuad Bajaber, Sherif Sakr, Omar Batarfi, Abdulrahman Altalhi, Radwa Elshawi, Ahmed Barnawi
- [11] Extreme-Scale Computer Architecture: Energy Efficiency from the Ground Up, Josep Torrellas