GyG

Get your Grocery AG

Our Big Data Architecture | Competitive advantage in the 21st century

Niels Humbeck (CDE)

21.03.2021 Köln

Agenda

1. GyG - Our Vision & Mission
2. Our Big Data Ambitions | GAP analysis
3. Core requirements of big data processing systems
4. Stream versus Batch processing of big data
5. GyG - Our big data architecture
6. One step ahead – shaping our future at GyG

GyG - Our Vision & Mission

Save peoples time by delivering the most efficient supply solution for groceries in the needed quality all over the world.

Our Big Data Ambitions:

Real time driven supply chain & sales

- Dynamic pricing
- Demand forecasting & optimization

World leading customer experience

- Product recommendation engines
- Interactive voice/ chat bots
- Personalized offers & advertisement
- Digital Store Assistant

Establishing multi-channel retailing

- Buy & collect
- Home delivery
- Localized product assortments

See Knight (2018)

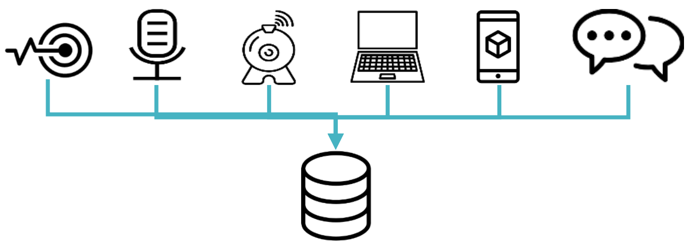
Our Big Data Ambitions | GAP analysis

Our Obsolete IT Infrastructure

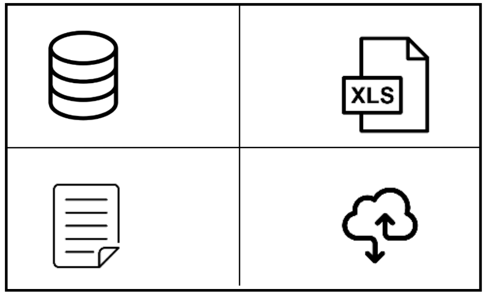
Our New IT Infrastructure

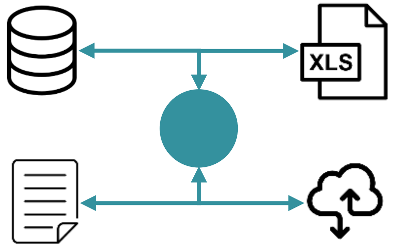
Multidimensional data ingestion

Costumer ID	Name	First Name	Age
50	Humbeck	Niels	29

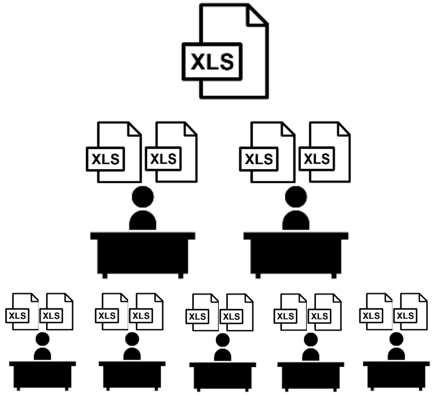


Connectivity & Latency





Increase analytic capabilities

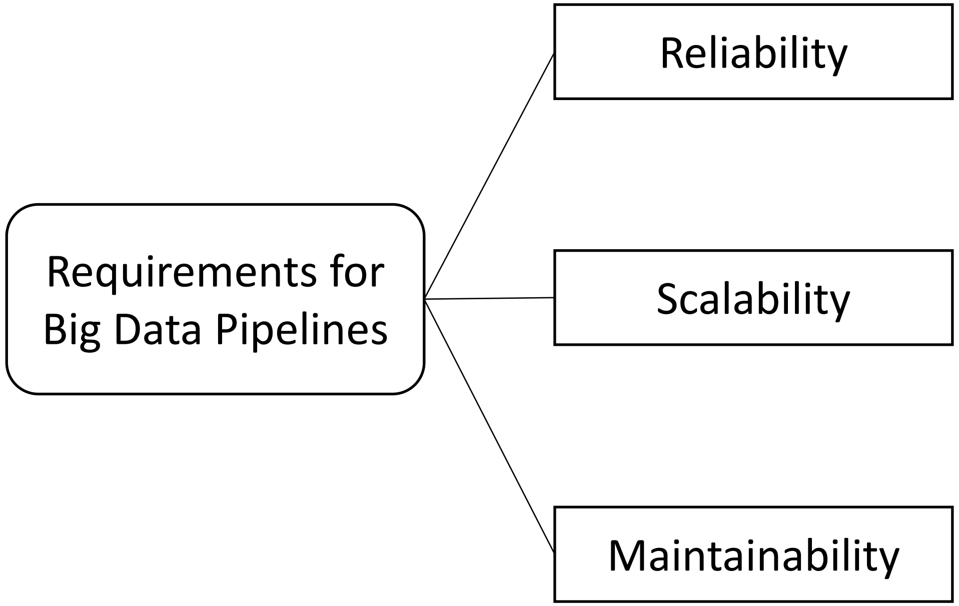


- Real Time analytics and visualization
- Automated reports
- Machine Learning

See Knight (2018)

Core requirements of big data processing systems

Reliability, scalability and maintainability



- The System should perform the correct function at the desired level of performance by preventing faults (hardware, software or human errors) causing failures & downtimes
- How to add resources to cope with increasing, data volume, traffic volume or complexity while keeping the performance?
- Maintain the functionalities in the most efficient way while being prepared for updates (operability, simplicity & evolvability).

Focus on:

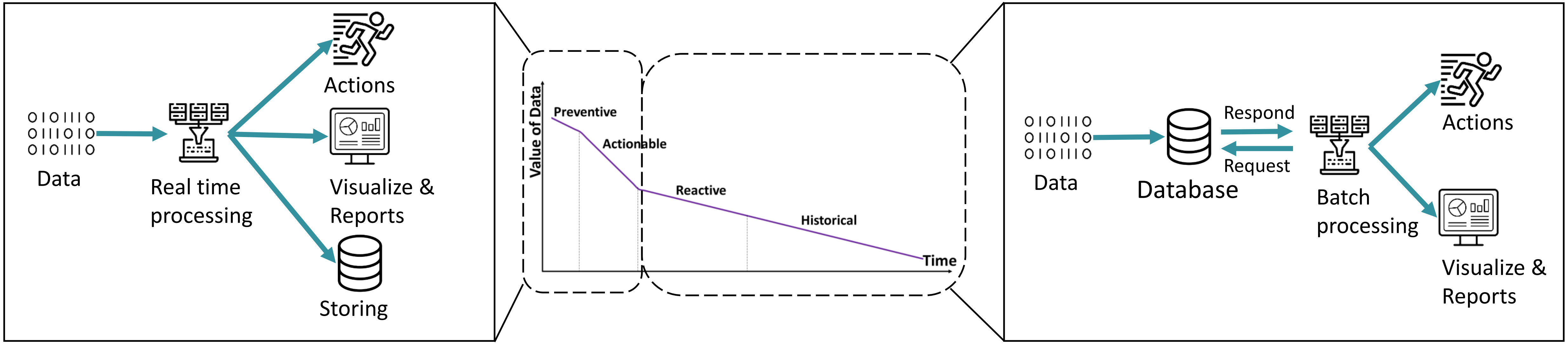
Optimized for availability, high throughput and low latency

See Kleppmann (2017), Marz (2015)

Stream versus Batch processing of big data

Stream processing:

Batch processing:



+
- Low latency
- Up to date data

−
- Expensive
- Higher complexity
- Normally less complex analysis

Use Cases:

- Realtime inventory tracking system
- Recommendation system

+
- Large batches of data
- Complex analytics & independency
- Efficient, cost effective

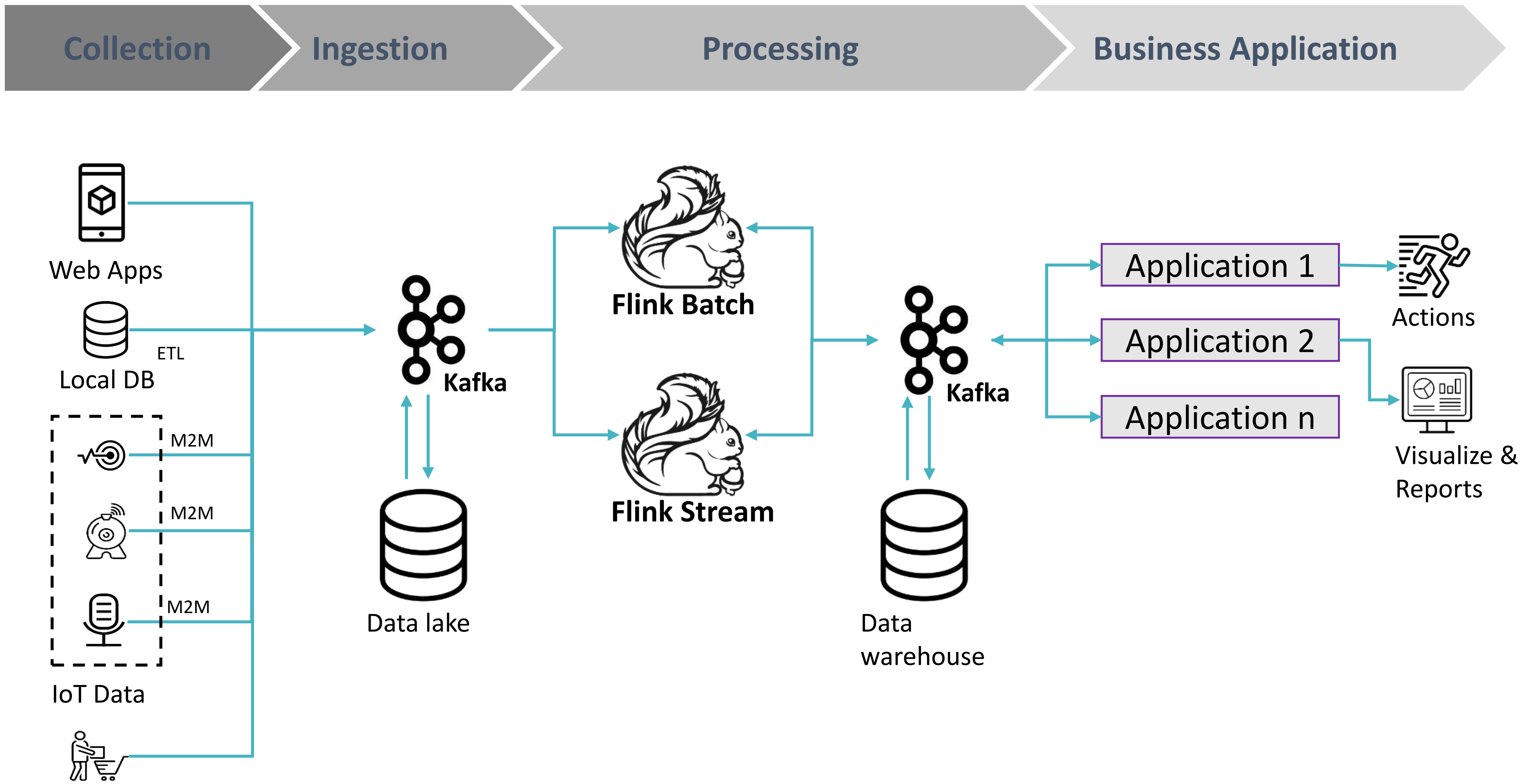
−
- High latency

Use Cases:

- Demand forecasting

See: Gualtiri (2016)

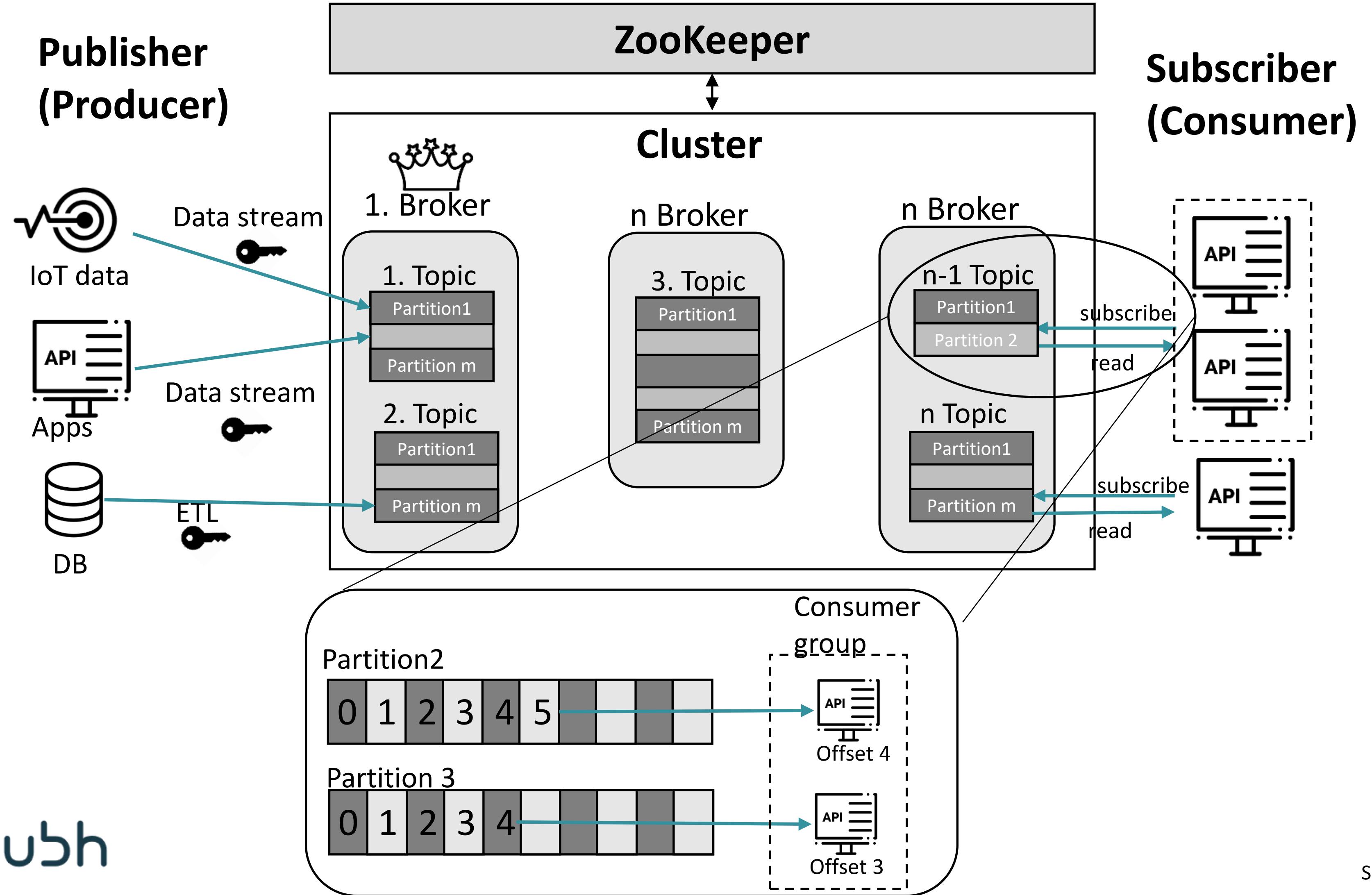
GyG - Our big data architecture



See Gupta (2020), Akanbi (2020), Nasiri (2019), Dendane (2019), Cheng (2018)

GyG - Our big data architecture

Kafka | Pub/Sub messaging system with distributed immutable commit log



Reliability guaranties:

- Guaranties the order in a partition
- At-least once messages guarantee

Scalability:

- Highly scalable (horizontal)

Maintainability:

- Multiple producers/ consumers
- Reduces integration complexity
- Highly configurable +retention
- Fault tolerant

+

- ETL & Big Data ingestion
- High Throughput
- Fairly low latency

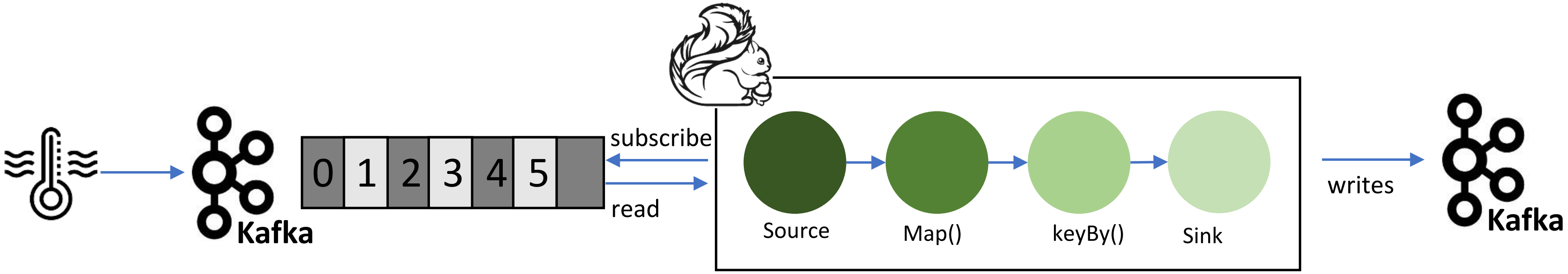
−

- Training
- Not for real low latency

See Müller-Kett (2020); Finematics(2019), Narkhede (2017)

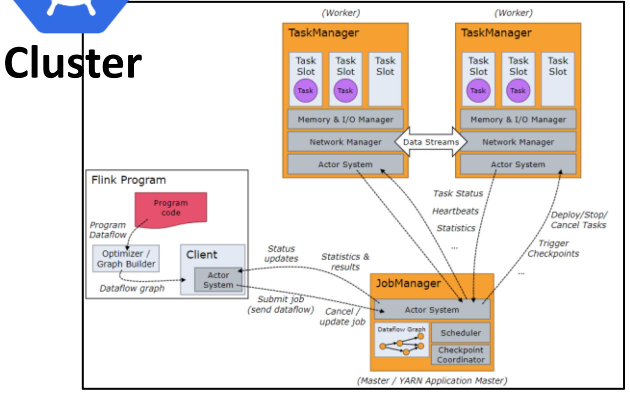
GyG - Our big data architecture

Flink | Distributed processing engine for stateful computations over (un-) & bounded data streams



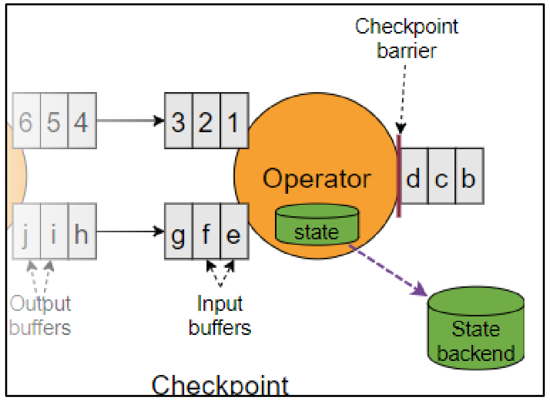
Cluster

Distributed



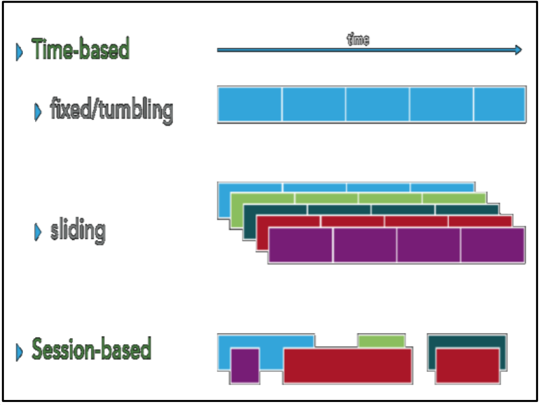
- Horizontal scaling
- High throughput

Stateful Processing



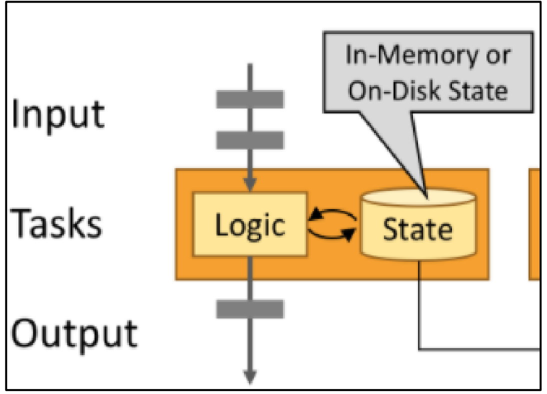
- Exactly once guarantee
- Fault tolerant

Timely Processing



- Enabling stream processing

In-Memory Processing

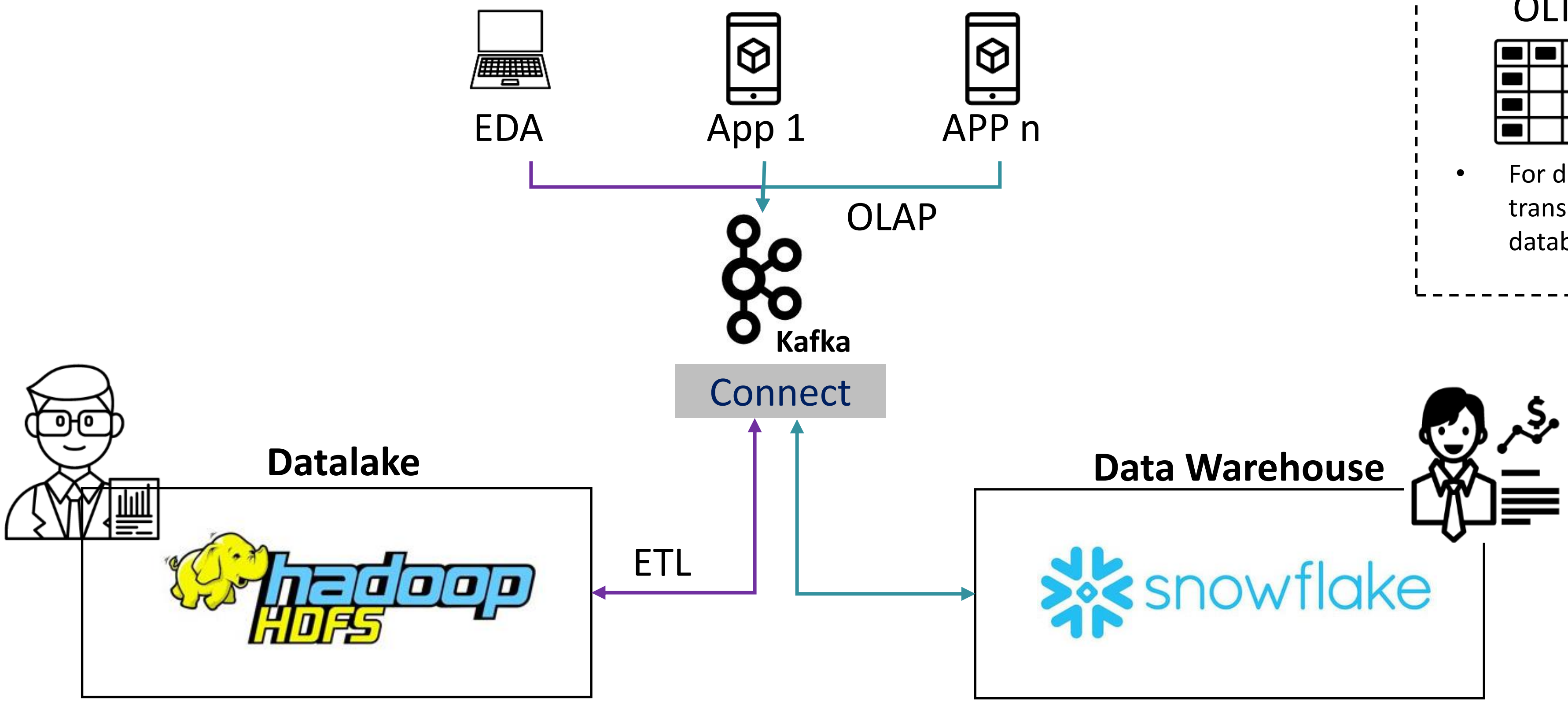


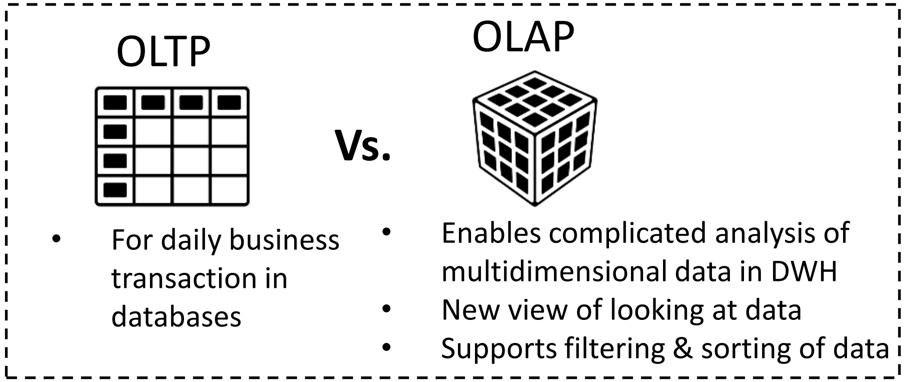
- Low latency

See Flink (1-3), Waski (2016), Sakr (2020)

GyG - Our big data architecture

Data storage & business aplication layer





OLTP
- For daily business transaction in databases

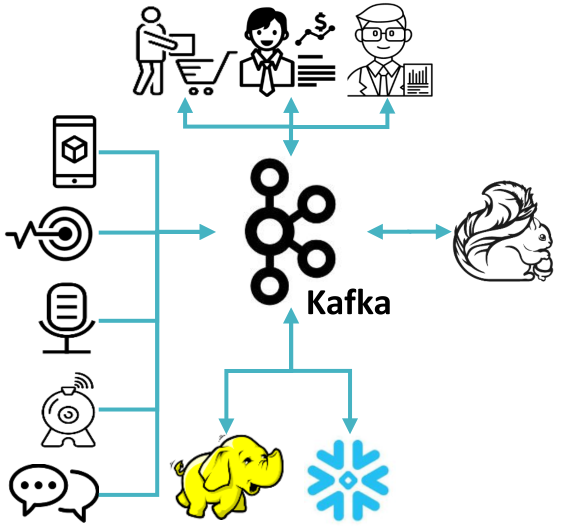
OLAP
- Enables complicated analysis of multidimensional data in DWH
- New view of looking at data
- Supports filtering & sorting of data

See Kidd (2020); Patil (2018), Chaudhuri (n.a.)

One step ahead – shaping our future at GyG

State of the art Big Data infrastructure

- Reliable, maintainable and scalable real time big data system
  - Kafka as our data backbone enabling multi data ingestion & low latency connectivity
  - Real time analysis with Flink
  - Next generation cloud data warehouse



✓

Personalized shopping experience & advertisement

Technology Trends:

- Facial recognition technology
- Social Media integration
  - Sentiment analysis (NLP)
  - Customer segmentation (ML)
- Interactive voice/ chat bots
- Block-Chain integration

Product Developments:

Personal store assistant (App)

- Product recommendation engines
- Virtual try-on functionalities
- Producti information

Personalized outdor advertisement

- Product recommendation engines
- Personalized outdoor advertisement

System Readiness: ✓

- Highy configurable data ingestion of multiple sources
- Processing, analysis & ML of various data in real time
- Store & access multidimensional data fast in the cloud

See Goasduff (2020), Reinhald (2020)

Q&A

Library (1/2)

- Akanbi, A., Masinde, M. (2020), A distributed Stream Processing Middleware Framework for Real-Time Analysis of Heterogeneous Data on Big Data Platform: Caes of Environmental Monitoring, Central University of Technology, South Africa
- Chaudhuri, S., Dayal, U. (n.a.), An Overview of Data Warehousing and OLAP Technology
- Cheng, C., Li, S., Ke, H. (2018), Analysis on the Status of Big Data Processing Framework, International Computers, Signals and Systems Conference
- Dendane, Y., Petrillo, F., Mcheick, H., Ben Ali, S. (2019), Quality model for evaluating and choosing a stream processing framework architecture, Universit du Qubec de Chicoutimi
- Finematics (2019), Apache Kafka Explained; https://finematics.com/apache-kafka-explained/, last access: 13.03.2021
- Flink 1 (2020), Flink Architecture, https://ci.apache.org/projects/flink/flink-docs-release-1.12/concepts/flink-architecture.html, last access 14.03.2021 at 11:31
- Flink 2(2020), Steteful Stream Processing, https://ci.apache.org/projects/flink/flink-docs-release-1.12/concepts/stateful-stream-processing.html, last access 14.03.2021 at 11:31
- Flink 3 (2020), What is Apache Flink?-Architecture, https://flink.apache.org/flink-architecture.html, last access 14.03.2021 at 11:31
- Foto 1ste Seite: https://thenounproject.com/photo/pattern-cubes-4dEanb/
- Goasduff, L. (2020), Gartner Top 10 Trends in Data and Analytics for 2020, https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-in-data-and-analytics-for-2020/, last access 21.03.2021 at 11:33
- Gualtiri, M., Curran, R. (2016). Perishable Insights – Stop wasting money on unactionable analytics. Forrester
- Gupta, S. (2020), Architecture for High-Throughput Low-Latency Big Data Pipeline on Cloud, https://towardsdatascience.com/scalable-efficient-big-data-analytics-machine-learning-pipeline-architecture-on-cloud-4d59efc092b5, last access 16.03.2021 at 21:14

Library (2/2)

- Kidd, C. (2020), Data Storage Explained: Data Lake vs Warehouse vs Database, https://www.bmc.com/blogs/data-lake-vs-data-warehouse-vs-database-whats-the-difference/, last access 14.03.2021 at 18:24
- Kleppmann, M. (2017). Designing data intensive applications: The big ideas behind reliable, scalable, and maintainable systems. Sebastopol, CA: O'Reilly
- Knight, T. (2018), Enabling new retail experiences with Big Data, https://www.youtube.com/watch?v=-HX-EI5uhsQ, last access 16.03.2021 as 21:01
- Marz, N., Warren, J. (2015), Big Data – Principles and best practises of scalable real time-time data systems, Manning Shelter Island
- Müller-Kett (2020); Course Book: Data Engineer – DLMDSEDE01, IUBH
- Nasiri, H., Nasehi, S., Goudarzi, M. (2019), Evaluation of distributed stream processing framewrks for IoT applications in Smart Cities, Journal of Big Data
- Narkhede, N. (2017), Exactly-Once Semantics Are Possible: Here's How Kafka Does it, https://www.confluent.io/blog/exactly-once-semantics-are-possible-heres-how-apache-kafka-does-it/, last access: 21.03.2021 at 12:27
- Patil, P. (2018), What is Explorative Data Analysis?, https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15, last access 14.03.2021 at 18:34
- Reinhald, M., Herhausen, D., Pahl, M, Wulf, J., 2020, Perspektiven für Face-Recognition im Data-Driven-Marketing, Marketing review St. Gallen
- Sakr, S. (2020), Big Data 2.0 Processing Systems-A Systems Overview, Springer, Institute of Computer Science, University of Tartu, Estonia
- EattleDataGuy (2020), What Are The Benefits Cloud Of Data Warehousing? And Should You Switch? https://medium.com/smb-lite/what-are-the-benefits-of-cloud-data-warehousing-a7322947a479, last access 21.03.2021 at 12:49
- Waski, A. (2016), Wiindowing data in Big Data Streams, Spark, Flink Kafka, Akka; https://softwaremill.com/windowing-in-big-data-streams-spark-flink-kafka-akka/, last access 14.03.2021 at 11:31

Sources | Icons (1/2)

- https://thenounproject.com/term/smart-home/1832362/
- https://thenounproject.com/search/?q=control+system&i=3340124
- https://thenounproject.com/search/?q=Protection&i=396887
- https://thenounproject.com/search/?q=utilities&i=2629112
- https://thenounproject.com/search/?q=robotic+cleaner&i=3307576
- https://thenounproject.com/term/machine-code/1706949/
- https://thenounproject.com/term/sensor/93304/
- https://thenounproject.com/term/web-camera/3409539/
- https://thenounproject.com/term/voice/3747767/
- https://thenounproject.com/term/antenna/1905220/
- https://thenounproject.com/search/?q=database&i=2321456
- https://thenounproject.com/term/action/1396548/
- https://thenounproject.com/search/?q=visualization&i=3060300
- https://thenounproject.com/search/?q=key&i=2474274
- https://thenounproject.com/search/?q=data+scientist&i=3463591

Sources | Icons (2/2)

- https://thenounproject.com/search/?q=businessman&i=3346861
- https://thenounproject.com/search/?q=app&i=1860579
- https://thenounproject.com/search/?q=Laptop&i=3029683
- https://thenounproject.com/term/table/250445/
- https://thenounproject.com/term/cube/1986590/
- https://thenounproject.com/search/?q=chat&i=2644028
- https://thenounproject.com/search/?q=excel&i=3267693
- https://thenounproject.com/search/?q=vision&i=1852050
- https://thenounproject.com/search/?q=sand+clock&i=3741831
- https://thenounproject.com/search/?q=check&i=1438093
- https://thenounproject.com/search/?q=mission&i=3405804
- https://thenounproject.com/term/advertisment/3014740/