Advanced Database Systems for Big Data - Challenges

Introduction to Stream Processing

Vera Goebel

Thomas Plagemann

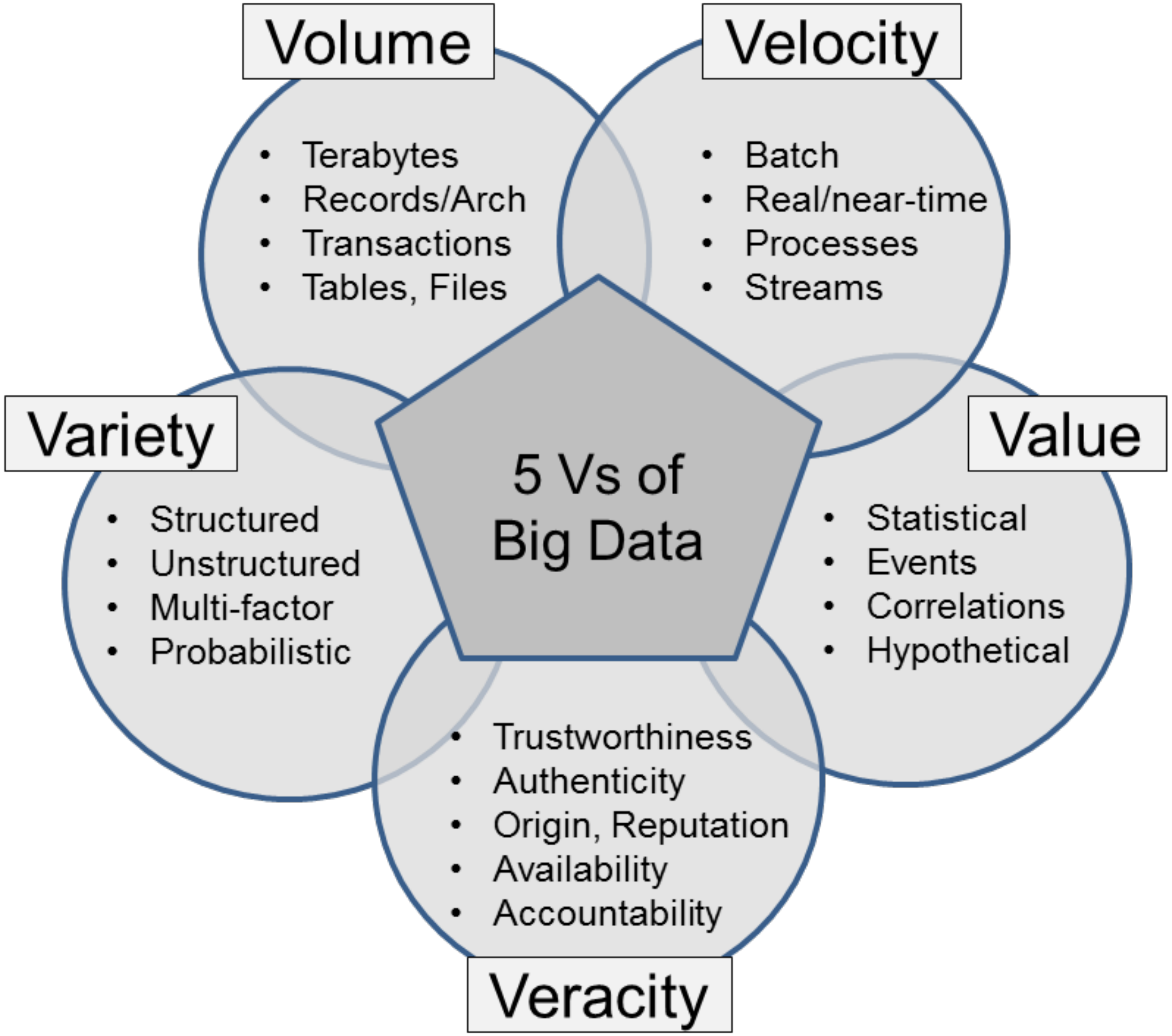
Report Series on Database Research

- Leading (30+) DB researchers & professors
- 10th meeting: 1989, 1990, 1995, 1996, 1998, 2003, 2008, 2013, 2018, 2023, …
- 3 newest reports focus on AI/ML & Big Data:
- The Cambridge Report on Database Research, A Ailamaki, S Madden, D Abadi, G Alonso, S Amer-Yahia, M Balazinska, PA Bernstein…, April 2025, arXiv preprint arXiv:2504.11259
- Abadi, D., et al.: The Seattle Report on Database Research, Communications of the ACM, August 2022, Vol. 65, No. 8, pp. 72-79
- Abadi, D., et al.: The Beckman Report on Database Research, Communications of the ACM, February 2016, Vol. 59, No. 2, pp. 92-99

Big Data – Defining Challenge [Beckman Report 2016]

- Distribution
- Integration
- Heterogeneity
- In-memory processing
- Large-scale systems
- Data analysis
- …

Big Data – Main Characteristics



Volume

- Terabytes
- Records/Arch
- Transactions
- Tables, Files

Velocity

- Batch
- Real/near-time
- Processes
- Streams

Variety

- Structured
- Unstructured
- Multi-factor
- Probabilistic

5 Vs of Big Data

Value

- Statistical
- Events
- Correlations
- Hypothetical

Veracity

- Trustworthiness
- Authenticity
- Origin, Reputation
- Availability
- Accountability

Big Data – 5 Related Challenges

- Scalable big/fast data infrastructures
- Coping with diversity in data management
- End-to-end data-to-knowledge pipeline
- Cloud services
- Role of people in data life cycle

Scalable Big/Fast Data Infrastructures

- Parallel and distributed processing
- Query processing and optimization
- New hardware
- Cost-efficient storage
- High-speed data streams
- Late-bound schemas
- Consistency
- Metrics and benchmarks

Diversity in Data Management

- No one-size-fits-all
- Cross-platform integration
- Programming models
- Data processing workflows

End-to-End Processing of Data

- Data-to-knowledge pipeline
- Tool diversity
- Tool customizability
- Open source
- Understanding data
- Knowledge bases

Cloud Services

- IaaS, PaaS, SaaS
- Elasticity (SLA)
- Data replication
- System administration and tuning
- Multitenancy (VMs)
- Data sharing
- Hybrid clouds

Roles of Humans in the Data Life Cycle

- Data producers
- Data curators (crowdsourcing)
- Data consumers
- Online communities

DB Research - New Challenges

[Beckman Report 2016] -> [Seattle Report 2022] - 1

- Correctly identified Big Data as a big theme
  - now morphed into data science, which poses big challenges
- Missed the AI/ML trend
- Promoted five directions
  - scalable data infrastructure
  - diversity in data management
  - end-to-end processing and understanding of data
  - cloud services
  - roles of humans in the data life cycle
- Made good progress, also branched out
  - e.g., into AI/ML

DB Research - New Challenges

[Beckman Report 2016] -> [Seattle Report 2022] - 2

- Beckman predicted the rise of a data-driven world
- Correctly observed that this gives us unprecedented opportunities & challenges:
  - Increasing amount and use of personal data -> data governance, ethical and fair use of data
  - Managed cloud data systems -> serverless systems, data lakes, ETL (extract, transform, load) jobs
  - Industrial Internet-of-Things (IoT)
  - Significant changes in hardware, esp. for ML/Deep Learning: FPGAs, GPUs, ASICs, …
- All of these have been true, but there are deep concerns that DBS community have failed to exploit this wealth of opportunities
  - while other communities have moved much faster.

DBS and Data Science ?

- Data Science: combines data cleaning and transformation, statistical analysis, data visualization, and ML techniques.
- Data Science [NSF CISE 2017]: “the processes and systems that enable the extraction of knowledge or insights from data in various forms, either structured or unstructured.”
- DBS technology plays a major role in Data Science: pipeline from raw input data to insights that requires use of data cleaning and transformation, data analytic techniques, and data visualization.
  - Data to insights pipeline
  - Data context and provenance
  - Data exploration at scale and data profiling
  - Declarative programming
  - Metadata management

Data Governance

- Data use policy -> GDPR, auditing
- Data privacy, e.g. differential privacy
- Ethical data science -> responsible data management

Cloud Services

- Serverless data services
- Disaggregation
- Multitenancy
- Edge and cloud
- Hybrid cloud and multi-cloud
- Auto-tuning
- SaaS cloud DB applications

Database Engines

- Heterogeneous computing
- Distributed transactions
- Data lakes
- Approximation in query answering
- Machine Learning (ML) workloads
- ML for reimaging data platform components
- Benchmarking and reproducability

DB Research - New Challenges

[Seatle Report 2022] -> [Cambridge Report 2025]

- Core Data Systems
  - Big Data everywhere, Cloud-based data systems, emerging new hardware, scalability, usability
- Human Centric Systems and Data Science
  - Big Data everywhere, data governance, NL-based querying & analysis interfaces
- ML and AI for Data Systems
  - AI/ML everywhere -> generative AI, LLMs, …
- Responsible Data Management

Core Data Systems

- Big Data everywhere
  -> usability, DB at massive scales
- Cloud-native architectures
- Disaggregated storage and compute
  -> high degree of scalability and flexibility
- Evolving hardware landscape
  -> resource-hungry AI, spec. AI accelerators

Human Centric Systems & Data Science

- Data sharing and collaboration
  -> break down data silos/lakes, enable cross-organizational analytics
- Privacy, governance, query processing across distributed datasets
- End-to-end data pipeline and workflow systems
  -> data discovery, explanations, preparation, integration and cleaning, metadata and log mgnt., versioning, analysis and visualization

ML and AI for Data Systems

- Query optimization, cost models based on ML
- Cardinality estimation, high-dimensional correlations in data distributions
- Reinforcement learning to improve physical data organization, predictive I/O
- Cloud resource management
- ML models for serverless VM management

Responsible Data Management

- Prevalence of AI models for interpreting data and making complex decisions
- Integrating data management research into responsible AI
- Decisions made during data collection and preparation impact accuracy, fairness, robustness, interpretablity, legal compliance of AI systems

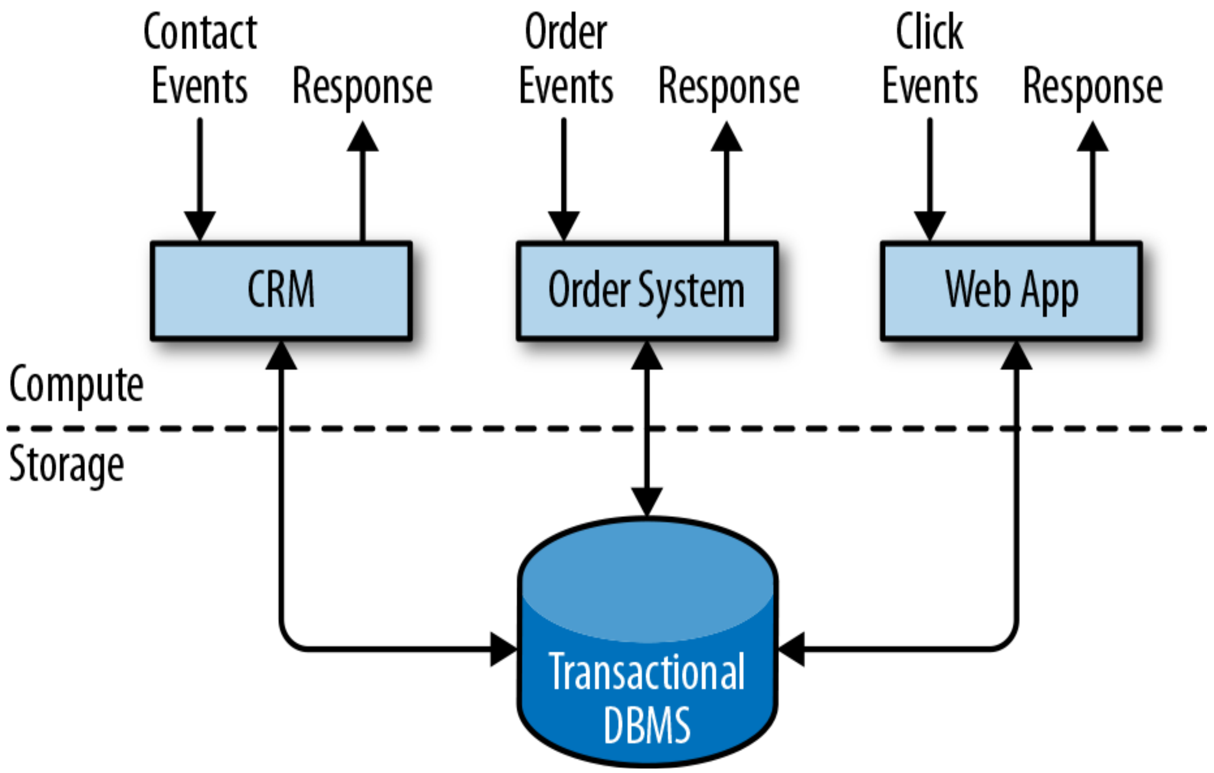
Content Overview

- Data Stream Processing (Part 1)
- Data Stream Processing (Part 2)
- Distributed Database Systems
- Heterogeneous Multi-Database Systems
- Web & XML Data Management
- Knowledge Discovery & Data Warehouses
- Machine Learning in Medicine *
- Scalable & Cloud Data Management
- Performance Analysis & Large DBS *

*Guest Lectures

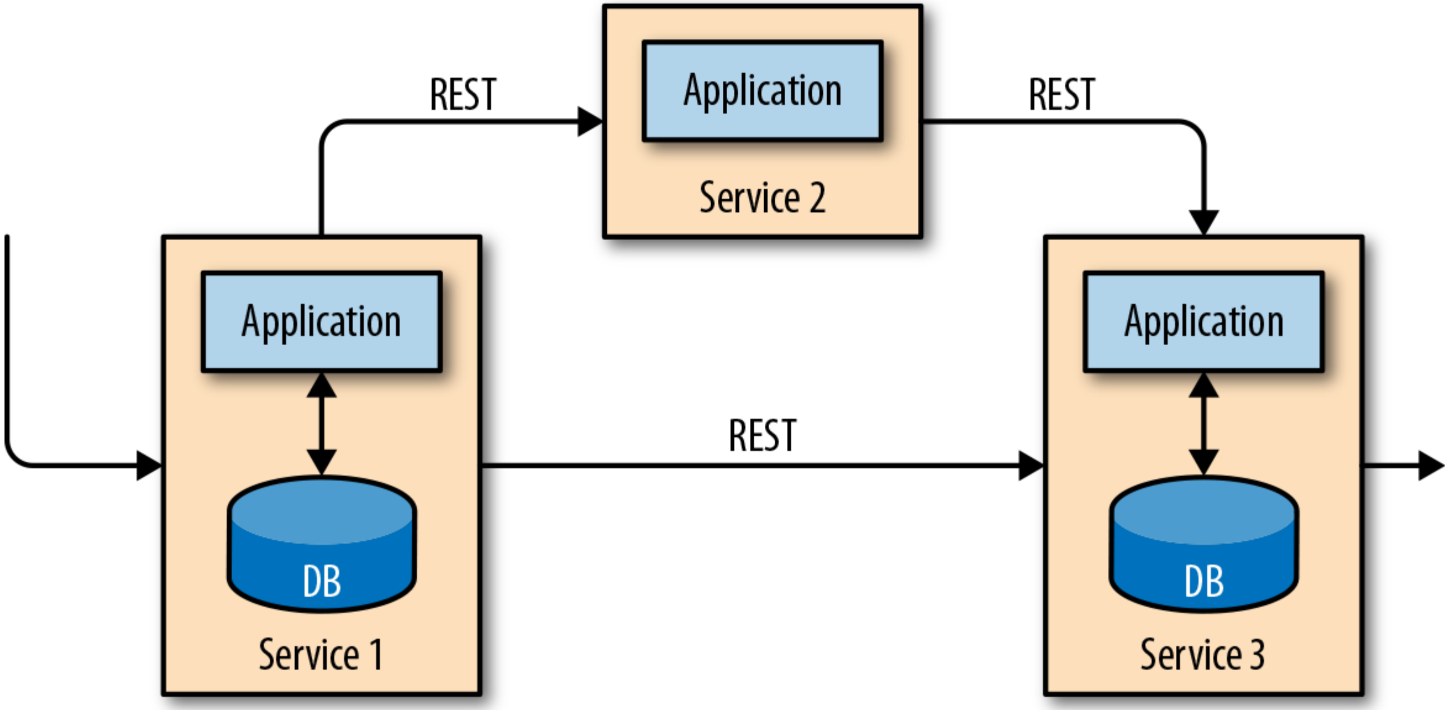
From traditional DBMS to heterogenous data processing infrastructure

Transactional Processing - Traditional



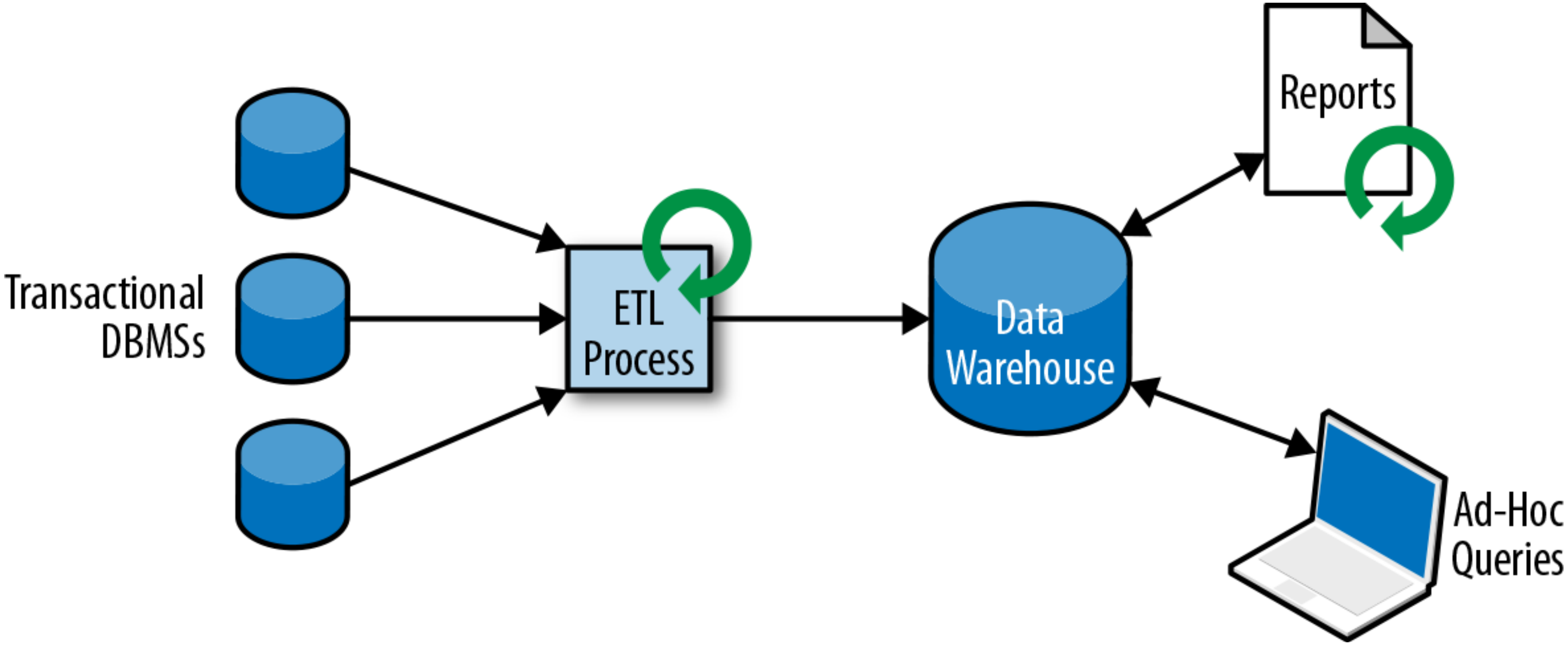
[Source: Stream Processing with Apache Flink, O´Reilly]

Transactional Processing – Micro Services



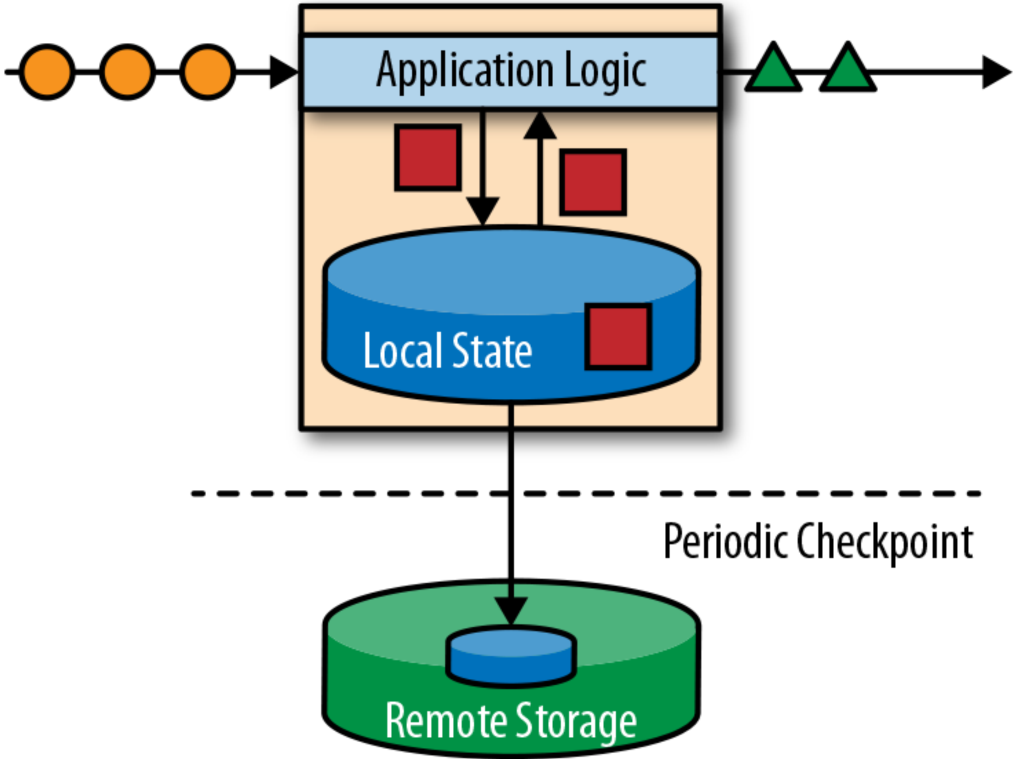
[Source: Stream Processing with Apache Flink, O´Reilly]

Analytical Processing – Data Warehouse



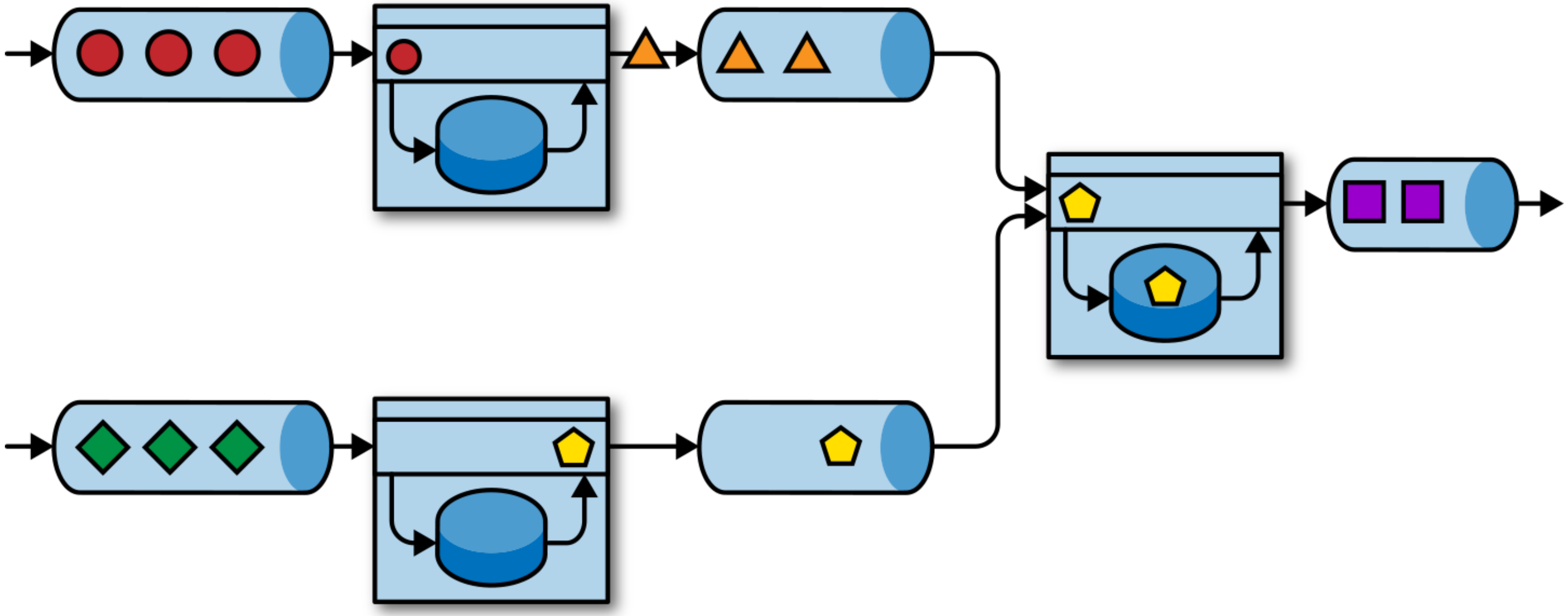
Source: Hueske, F., Kalavri, V.: Stream Processing with Apache Flink, O´Reilly, 2019

Analytical Processing – Stateful Stream Processing



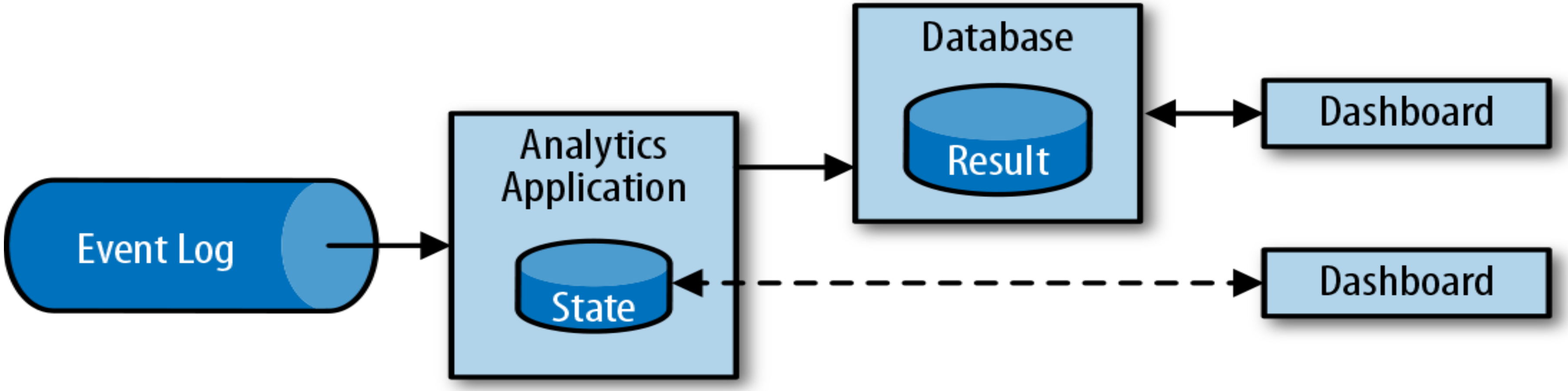
Source: Hueske, F., Kalavri, V.: Stream Processing with Apache Flink, O´Reilly, 2019

Analytical Processing – Event Driven Architecture



Source: Hueske, F., Kalavri, V.: Stream Processing with Apache Flink, O´Reilly, 2019

Streaming Analytics



Source: Hueske, F., Kalavri, V.: Stream Processing with Apache Flink, O´Reilly, 2019

Looking for Master thesis topics?

- We are part of the DKM group at Ifi:
  https://www.mn.uio.no/ifi/english/research/groups/dkm/
- Respire project:
  Responsible Explainable Machine Learning for Sleep-related Respiratory Disorders
  https://www.mn.uio.no/ifi/english/research/projects/respire/index.html
- Parrot project:
  Privacy Engineering for Real-Time Analytics in Human-Centered Internet-of-Things
  https://www.mn.uio.no/ifi/english/research/projects/parrot/index.html

- Contact: plageman@ifi.uio.no