Data lakes and data warehouses

Tools to keep your data cold, warm or hot

Where should I store my data?

This overview aims to contrast weakly structured data lakes with traditional relational data warehouses. Contemporary information systems are placed on a map ranging from cold, infrequently accessed data to warm, regularly queried data up to hot, business-critical data that is always kept in memory. Many modern solutions are built to span multiple temperature zones and attempt to offer a structured relational interface that abstracts from the underlying heterogeneity.

Apache Flink

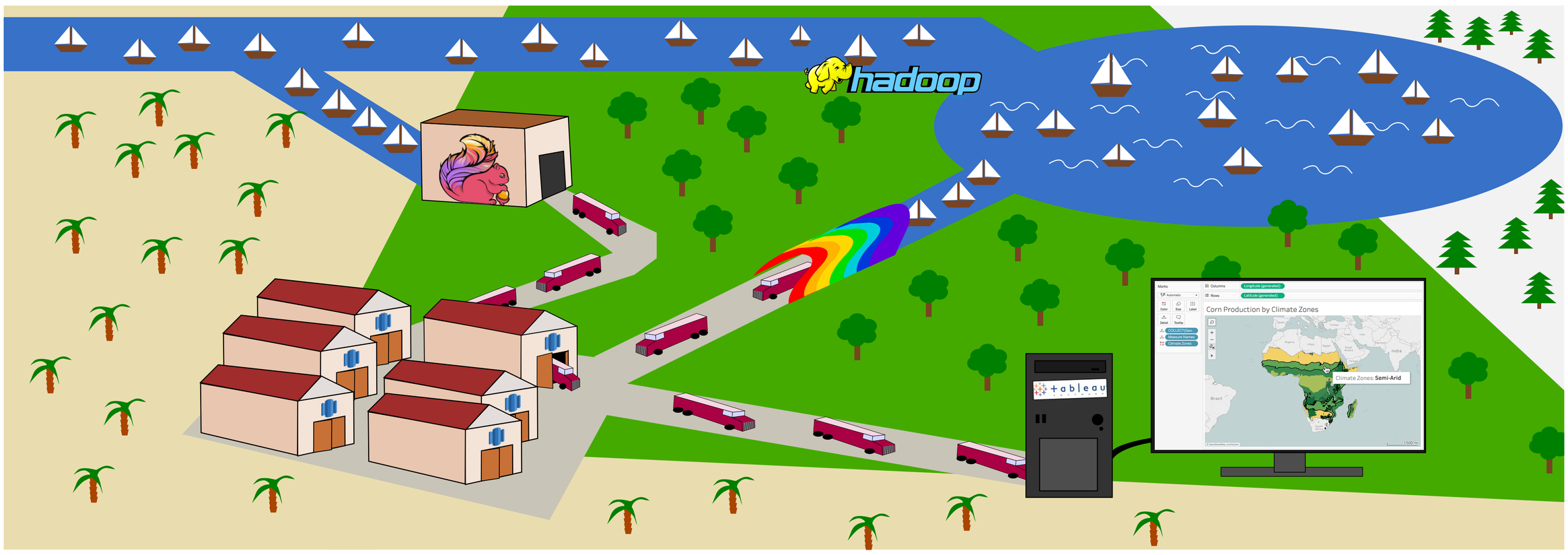
This stream processing framework is commonly used for real-time analysis and data warehouse preprocessing.

Data Lakes

Enterprises will store all collected and produced data from database entries to text, images and video in massive data lakes that require little to no preprocessing and come at a cheaper price point than data warehouses.

Apache Hadoop

This commodity-cluster optimized pairing of HDFS and MapReduce is a widely adopted base for data mining applications.



Data Warehouses

Relational data warehouses integrate data from multiple sources to facilitate online analytical processing (OLAP). This enables data-driven decision making.

Amazon Redshift (Spectrum)

Redshift is a relational data warehouse cloud service based on PostgreSQL. Redshift Spectrum allows to query Amazon S3 data lakes with SQL to simplify the analysis of weakly structured data sets.

Tableau

Even in the times of automated decision making, visualization software is still necessary for exploration of unknown data sets and support of complex strategic decisions.

References

Martin Grund. 2018. Exabytes for Breakfast. Video. (January 9, 2018). Retrieved January 30, 2018 from https://www.tele-task.de/lecture/video/6627/
Lennart Heuckendorf. 2017. The science behind Visual Analytics. Video. (November 14, 2017). Retrieved January 30, 2018 from https://www.tele-task.de/lecture/video/6477/
Fabian Hüske. 2017. Modern Stream Production with Apache Flink. Video. (November 28, 2017). Retrieved January 30, 2018 from https://www.tele-task.de/lecture/video/6547/
Brian Stein and Alan Morrison. 2014. The enterprise data lake: Better integration and deeper analytics. Technology Forecast 2014-1. PwC Center for Technology and Innovation.

Author

Julius Lischeid
Bachelor student
julius.lischeid@student.hpi.de

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam