High-Performance-Datenanalysen im Exabytebereich
Ein Vergleich von Batch und Stream Processing

Zusammenfassung

Unternehmen vieler Branchen generieren bei ihrer täglichen Arbeit Petabytes an Daten. Zwar ist die Speicherung selbst solcher Datenmengen finanzierbar geworden, jedoch verhinderten die beträchtlichen Laufzeiten aktueller Analysealgorithmen bisher einen die Serverkosten rechtfertigenden Erkenntnisgewinn aus diesen Daten.

Big-Data-Unternehmen und Open-Source-Communities arbeiten aktuell an verschiedenartigen Lösungsansätzen. Dieses Plakat diskutiert die Vor- und Nachteile der Methoden Stream und Batch Processing an den State-of-the-art-Tools Apache Flink und Amazon Redshift Spectrum.

Apache Flink

Ein Open-Source-Framework für verteilte Verarbeitung von Datenströmen.

Vorteile

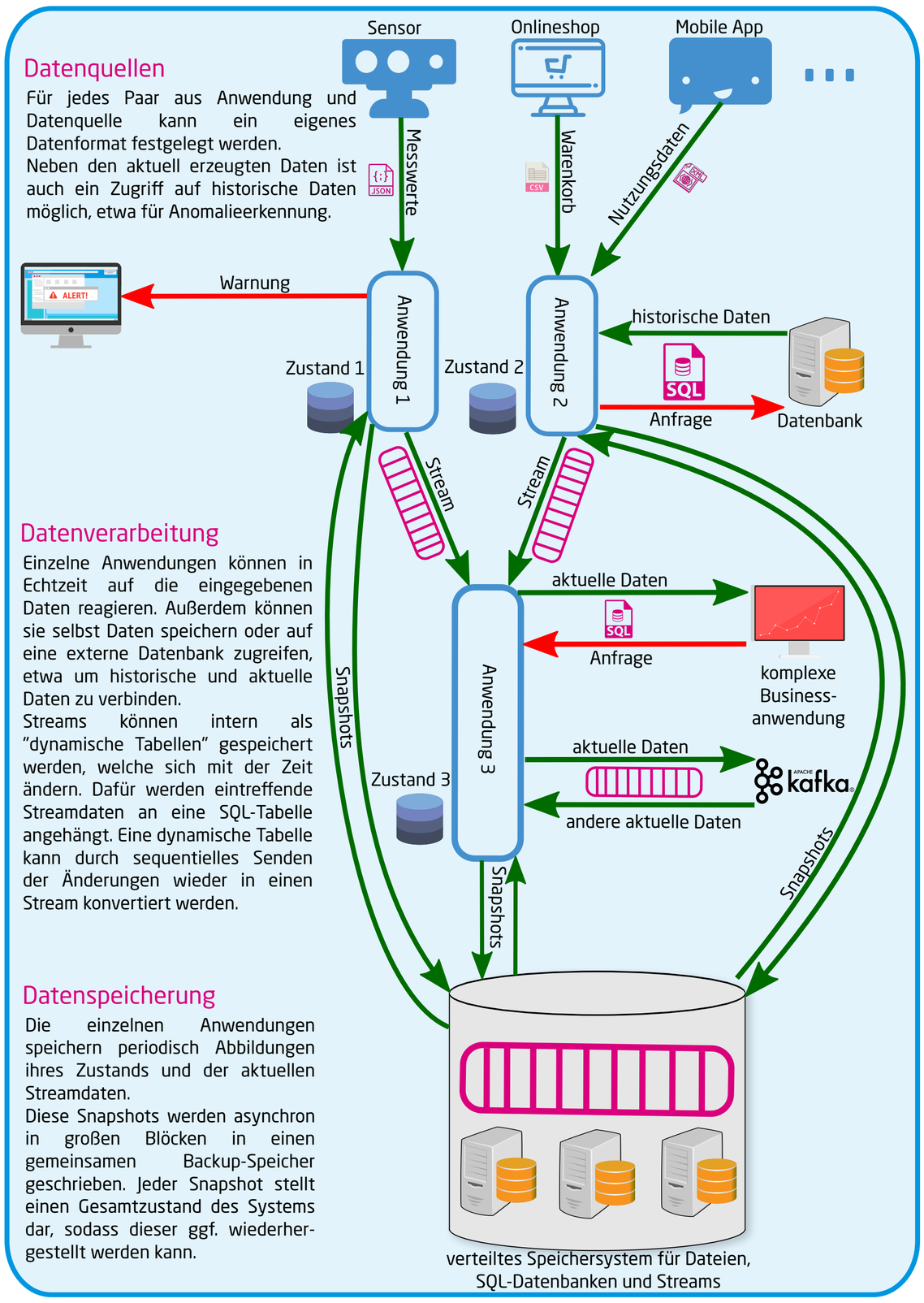
- Batch und Stream Processing in einem System
- unterstützt viele Formate, Bibliotheken und Standards, Integration in größeres Big Data Framework möglich
- Datenhoheit des Anwenders bleibt durch Verwendung eigener Server erhalten
- Fehlertoleranz durch Snapshot-System
- stark nebenläufig und gut skalierbar
- Stream muss nicht komplett gespeichert werden, wenn die Speicherung verschiedener Verarbeitungsergebnisse ausreicht
- 4 Abstraktionsebenen für verschiedene Anwendungen
- kostenfreie Nutzung, Quelloffenheit

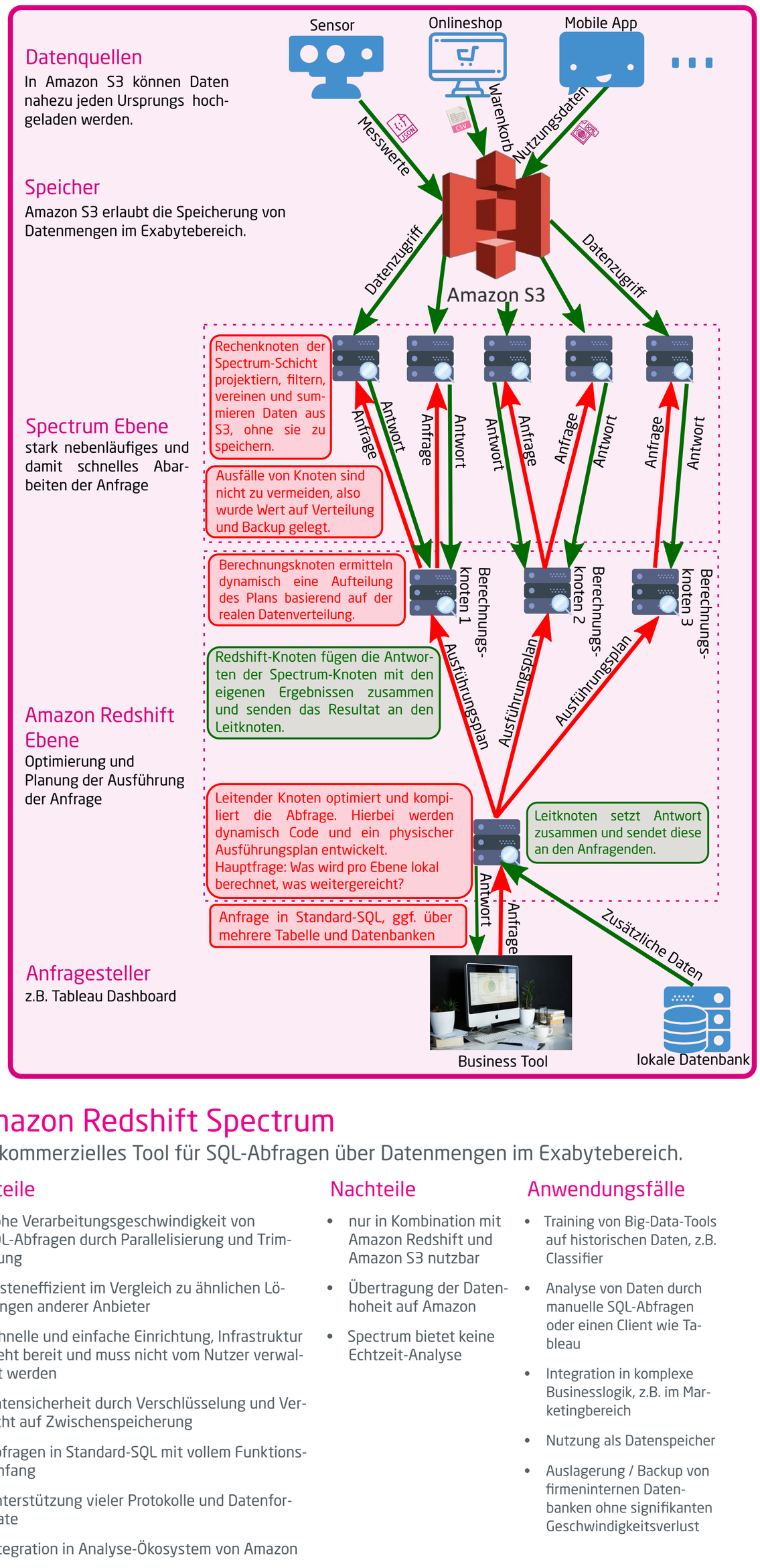
Nachteile

- aufwendige Einrichtung, vieles ist selbst zu Implementieren
- tatsächliche Arbeitsgeschwindigkeit hängt von vielen Faktoren ab
- Beschränkungen der Abfragemöglichkeiten für dynamische Abfragen
- optimiert für schnell veränderliche Daten und konstante Abfragen (Stream Processing), deswegen evtl. niedrigere Geschwindigkeit beim Batch Processing als andere Tools

Anwendungsfälle

- Kreditkartenbetrug live erkennen
- Marketing-Systeme für Online-Shops (Empfehlungen und Angebote in Echtzeit)
- Überwachung technischer Anlagen
- wissenschaftliche Experimente, welche große Datenmengen produzieren, bei denen nur einzelne Charakteristika relevant sind (z.B. Teilchenbeschleuniger)
- Analyse von Social Media in Realzeit





Amazon Redshift Spectrum

Ein kommerzielles Tool für SQL-Abfragen über Datenmengen im Exabytebereich.

Vorteile

- hohe Verarbeitungsgeschwindigkeit von SQL-Abfragen durch Parallelisierung und Trimmung
- kosteneffizient im Vergleich zu ähnlichen Lösungen anderer Anbieter
- schnelle und einfache Einrichtung, Infrastruktur steht bereit und muss nicht vom Nutzer verwaltet werden
- Datensicherheit durch Verschlüsselung und Verzicht auf Zwischenspeicherung
- Abfragen in Standard-SQL mit vollem Funktionsumfang
- Unterstützung vieler Protokolle und Datenformate
- Integration in Analyse-Ökosystem von Amazon
- Unterstützt Zugriff auf Dateien und Datenbanken

Nachteile

- nur in Kombination mit Amazon Redshift und Amazon S3 nutzbar
- Übertragung der Datenhoheit auf Amazon
- Spectrum bietet keine Echtzeit-Analyse

Anwendungsfälle

- Training von Big-Data-Tools auf historischen Daten, z.B. Classifier
- Analyse von Daten durch manuelle SQL-Abfragen oder einen Client wie Tableau
- Integration in komplexe Businesslogik, z.B. im Marketingbereich
- Nutzung als Datenspeicher
- Auslagerung / Backup von firmeninternen Datenbanken ohne signifikanten Geschwindigkeitsverlust

Quellen

Vorträge
Grund, Martin (2018). Exabytes for Breakfast. Vortrag in Ringvorlesung „Data Engineering in der Praxis", Hasso-Plattner-Institut Potsdam.

Hüske, Fabian (2017). Modern Stream Production with Apache Flink. Vortrag in Ringvorlesung „Data Engineering in der Praxis", Hasso-Plattner-Institut Potsdam.

Webseiten
http://flink.apache.org, Abruf am 29.01.2017 18:22

Bildquellen
Icons für Sensor, Online Shop, SQL: Nutzer Freepik auf www.flaticon.com

Icons für JSON, CSV, Server, Datenbank: Nutzer Smashicons auf www.flaticon.com

Logo Amazon S3: http://www.vmtocloud.com/wp-content/uploads/2017/04/S3.jpeg

Logo App: Nutzer SimpleIcon auf www.flaticon.com

Weitere Symbole und Bilder verwendet unter Creative Commons 0 Lizenz ohne Verpflichtung zur Namensnennung. Abrufdatum aller Links: 30.01.2018 17:51.

Autor

Eric Ackermann
Bachelor IT-Systems Engineering
1. Semester

Prof.-Dr.-Helmert-Str. 2-3
D-14482 Potsdam
E-Mail: eric.ackermann@student.hpi.de

Hasso Plattner Institut
Digital Engineering · Universität Potsdam