Big Data

Lotaire Tchamadeu Tiappi
Angewandte Informatik HAW
Hamburg, Deutschland
Lotaire.tchamadeutiappi@haw-hamburg.de

Agenda

- Big Data und die Paketen Lieferung
- Bisherige Arbeit
- Thema
- Forschungsfrage
- Motivationen
- Methoden
- Vorgehen
- Ergebnis
- Referenzen

Big Data und die Paketlieferung

- Big Data wird oft unter diesen 4 Wörtern definiert [1]
- Volume (Umfang der Daten):
  - 2 bis 3 Millionen Pakete/Tag
- Velocity (Geschwindigkeit, in der die Daten erfasst werden):
  - 10000-15000 Touren MB/s, item/s
- Variety (vielfältige Arten von Daten):
  - Strukturierte Daten : ankommende Paketinformationen in JSON Format
  - semi-strukturierte Daten : Mails von Lieferanten
  - Unstrukturierte Daten : Scanner Daten oder Sensor Daten
- Veracity (Wahrhaftigkeit der Daten)
  - In den meisten Fällen werden einige Transformationen erforderlich

Was ist Big Data, Thema, Bisherige Arbeit, Forschungsfrage, Motivationen, Methode, Vorgehen, Ergebnisse, Referenzen

Thema

- Streaming Szenario zur Unterstützung der Ankündigungen von Paketen im Logistikbereich

Was ist Big Data, Thema, Bisherige Arbeit, Forschungsfrage, Motivationen, Methode, Vorgehen, Ergebnisse, Referenzen

Bisherige Arbeit

- Grundseminar
  - Anwendungsspezifische proaktive Behandlungen von Prozessen im Bereich Big Data
- Grundprojekt
  - Integration von Big Data´s Technologie in Anwendungslandschaft
- Hauptprojekt
  - Streamings Prototyp im Bereich der Logistik

Was ist Big Data, Thema, Bisherige Arbeit, Forschungsfrage, Motivationen, Methode, Vorgehen, Ergebnisse, Referenzen

Motivationen

- Unterstützung der Logistik bei der Paketankündigung im Rahmen der Paket-Verfolgung. (echtzeitige Geolokalisierung von Paket-Daten)
- Kundenzufriedenheit
  - Kunden können die Ankündigungszeit aus der Prognose mehr vertrauen, indem sie selbst jeder Zeit verfolgen können, wo sich ihre Pakete gerade befinden.
- Die Logistik kommt schneller an ihre Daten als vorher
- Streaming Daten können in Echtzeit ausgewertet werden.
- Betrugserkennung und Paketverlust senken (momentan um 0, 03 %)

Forschungsfrage

Wie würde der Prototyp des Streaming Processing bei der Paketlieferung im Bereich der Logistik aussehen? Welche Parametrisierung soll bei der Auswahl der Technologien betrachtet werden und welche Konsequenzen ergeben sich für die Geschäftsstruktur?

Methoden

- Entwicklung eines Streaming Prototypen für Scanner Daten

Was ist Big Data, Thema, Bisherige Arbeit, Motivationen, Forschungsfrage, Methode, Vorgehen, Ergebnisse, Referenzen

Big Data Technologien

The Big Data technology stack is evolving rapidly

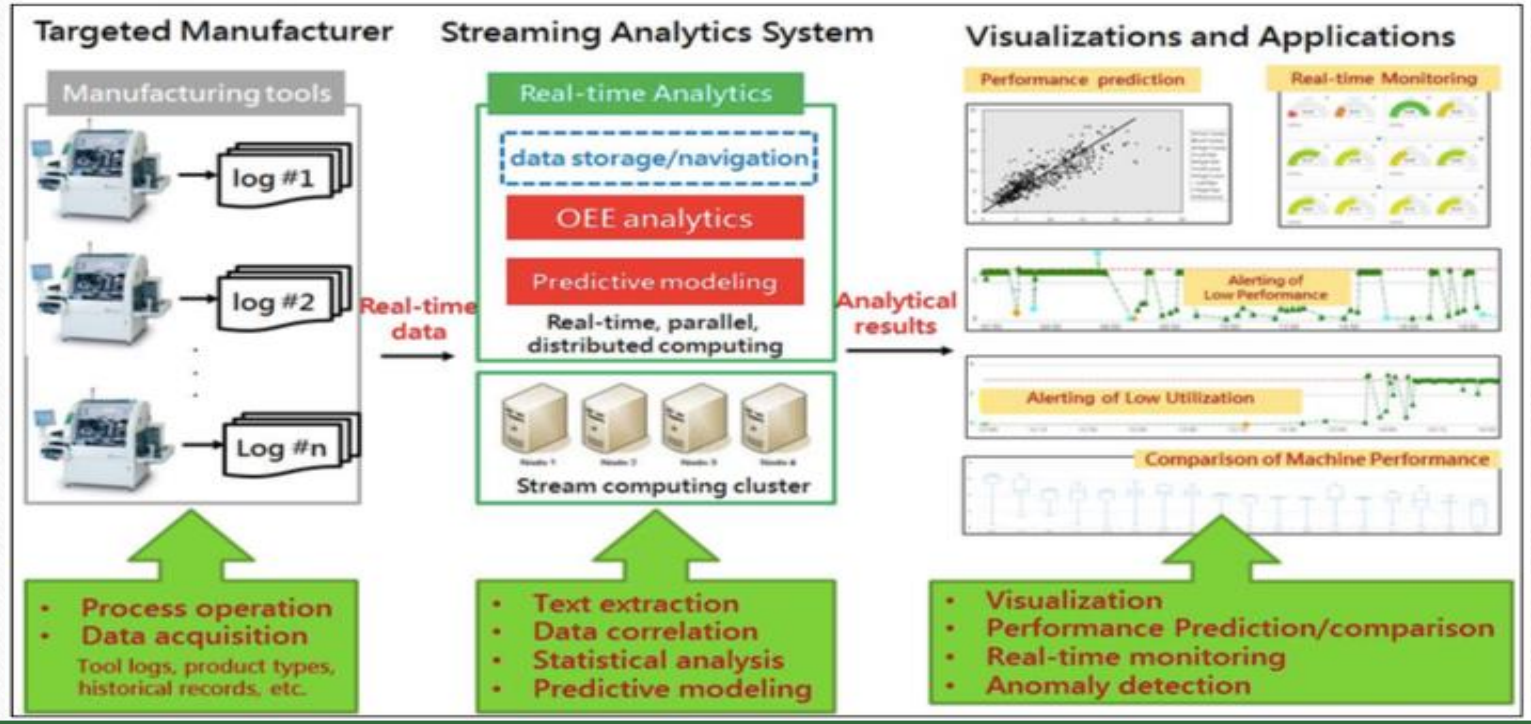
[2]

Was ist Big Data, Thema, Bisherige Arbeit, Motivationen, Forschungsfrage, Methode, Vorgehen, Ergebnisse, Referenzen

Example Streaming analytics processing in the manufacturing operation

Wir können dieses Exemplar unserem Konzept übertragen.

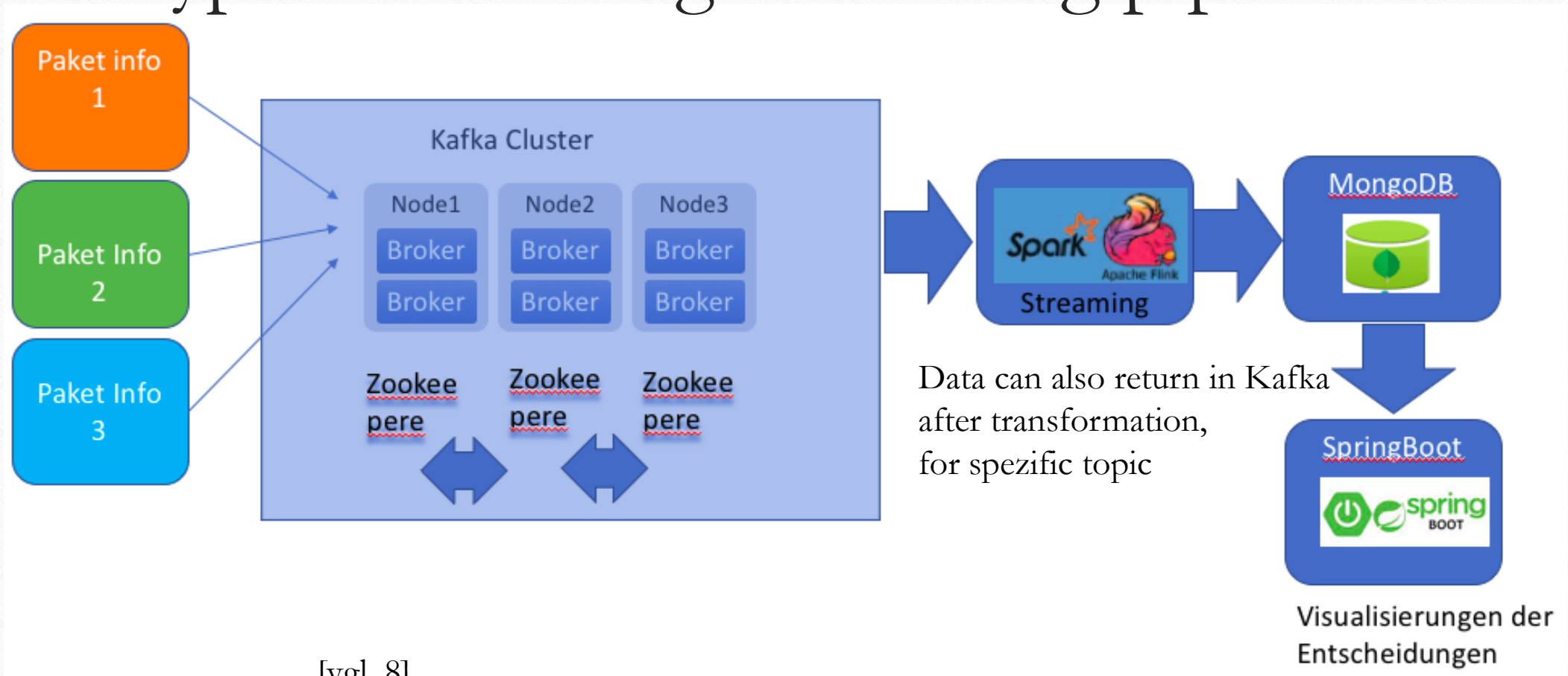
Für Akquisition von Daten werden wir es mit Hilfe von Apache Kafka machen



[3]

Was ist Big Data, Thema, Bisherige Arbeit, Motivationen, Forschungsfrage, Methode, Vorgehen, Ergebnisse, Referenzen

Typical streaming Processing pipeline



[vgl. 8]

Vorgehen

- Definition der KPI zur Beobachtung des Prototyps
- Festlegung der Streamings Metrik und KPI(Key Performance Indikator)[4,5,6]

Gesamtanlageneffektivität (GAE)

  - Leistungsfaktor (L)
  - Verfügbarkeitsfaktor (V)
  - Qualitätsfaktor (Q)

- GAE= L * V *Q

Leistung	Verfügbarkeit	Qualität	Parameter
Latenz: Bearbeitungszeit eine individuelle Mikrobündel aus wenigen Aufzeichnung. Erfordert Latenz im Bereich von Sekunden oder Millisekunden[9]	Gesamtlaufzeit der Komponenten	theoretische Bearbeitungszeit für ein effektives Mikrobündel	Anzahl der Broker in Cluster; Anzahl der File Zookeeper (Zookepeer Cluster)
Durchsatz:(10000-15000 Touren MB/s, item/s)	Gesamtausfallzeit	theoretische Bearbeitungszeit für ein tatsächliches Mikrobündel	Batch size
Analyse :Einfache Reaktionsfunktion, Aggregate und gleitende Metriken			Aneignungsstrategie
CPU Auslastung			Nachrichtenreplikation
Festplatten Auslastung			Verwendete Hardware; Threads for sending the data on the network and for disk I/O
Netzwerk Auslastung			Nachrichtgröße
Speicher Auslastung			

Was ist Big Data, Thema, Bisherige Arbeit, Motivationen, Forschungsfrage, Methode, Vorgehen, Ergebnisse, Referenzen

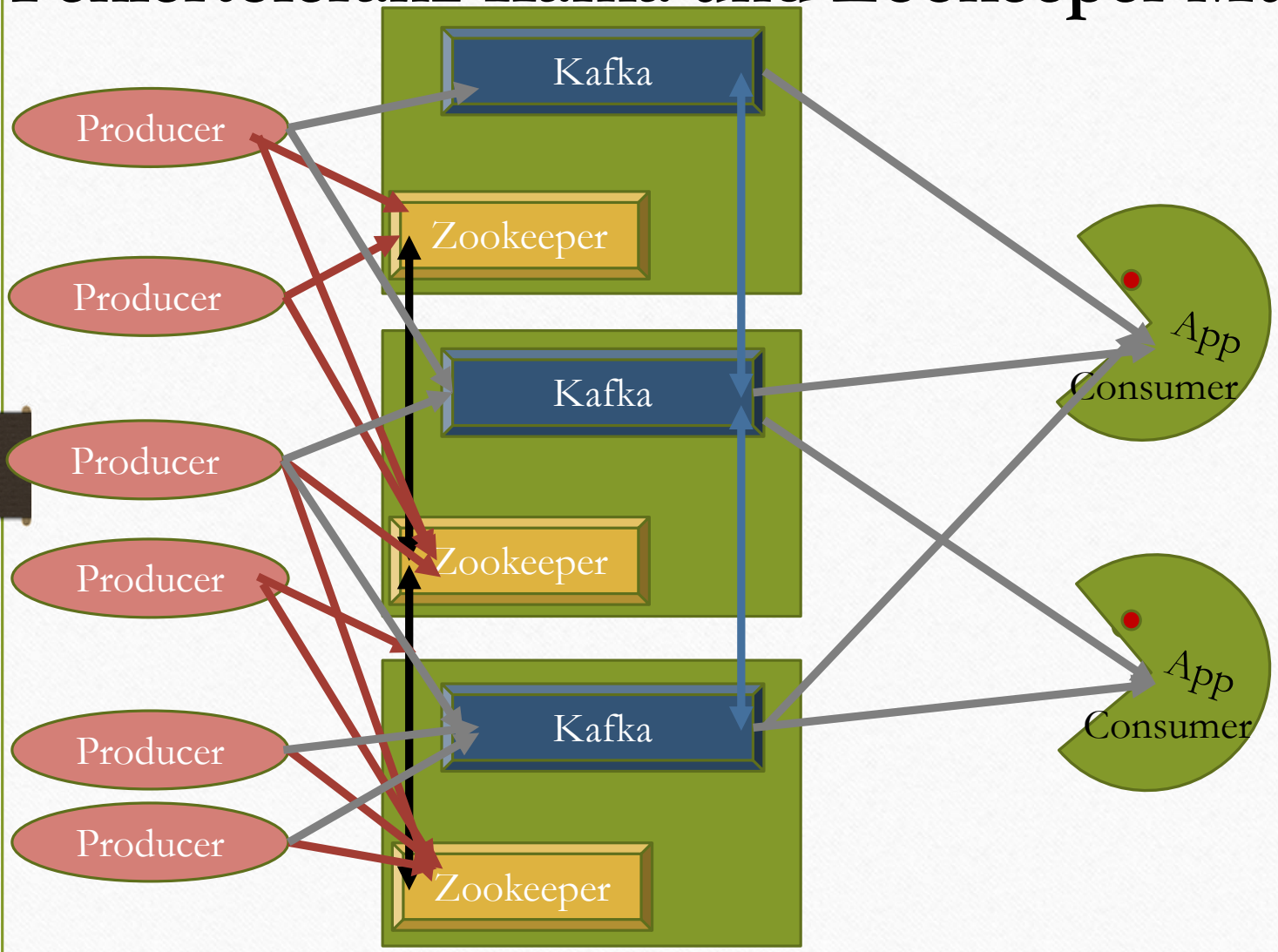
Vorgehen

- Ein vorläufiges Design für das neue System wird auf Basis von Apache Kafka erstellt.

Dann wird eine alternative Lösung auch im Ansatz gebracht (zum Beispiel Apache Flume)

- Producer API für Scanner-Daten implementieren
- Consumer-API für Daten-Verbrauchen festlegen
- Wir werden danach sukzessive Apache Kafka Streaming, Spark Streaming und Flink Streaming einbinden und die Fehlertoleranz beobachten.
    - (Knoten werden von Cluster ausgeschaltet)
- Der Streaming mit der besten Fehlertoleranz wird ausgewählt.

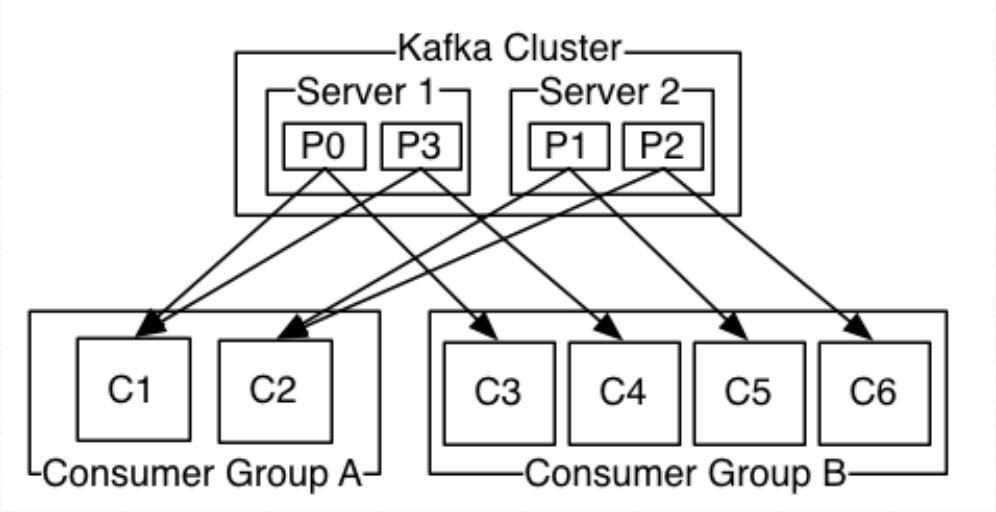
Fehlertoleranz Kafka und Zookeeper Multi Node Cluster



Bei Ausfall eines Knoten wird seine Trafik von anderen Knoten übernommen.

Ein vorläufiges Design

- Kafka Cluster einrichten
  - Broker einrichten
  - Topic definieren
  - Partition definieren



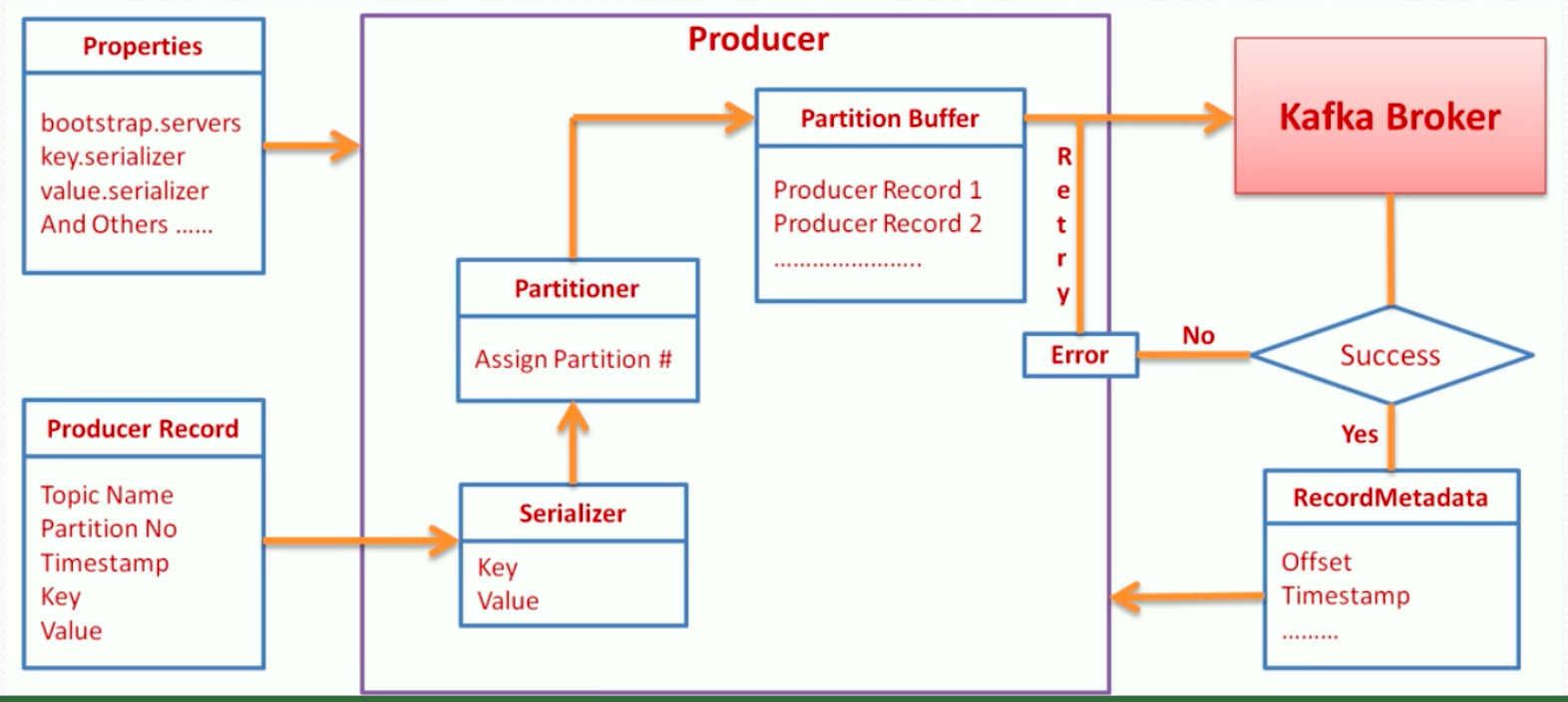
[7]

Implementierung von Producer API

- Einbindung von Scanner-Daten mit Kafka Server.
    - Wir bekommen von jeden gescannten Paketen die Geodaten-information

Vgl.[8]

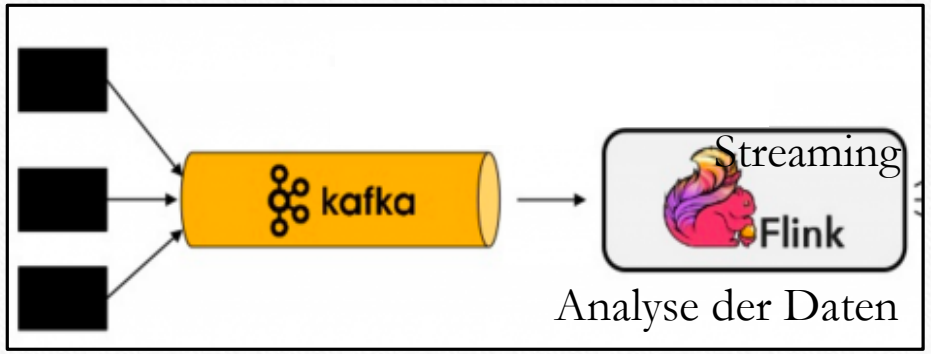
Producer Workflow

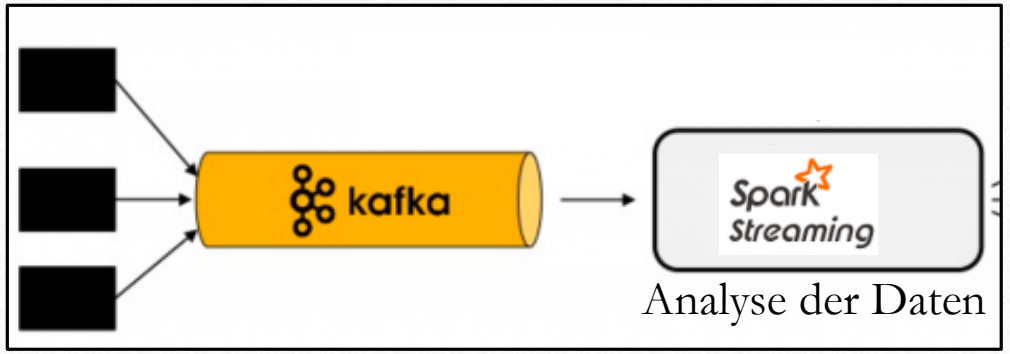


Was ist Big Data, Thema, Bisherige Arbeit, Motivationen, Forschungsfrage, Methode, Vorgehen, Ergebnisse, Referenzen

Streaming Frameworks einbinden

Wir möchten die Fehlertoleranz bei den beiden Streaming testen
und uns für eine entscheiden.





Was ist Big Data, Thema, Bisherige Arbeit, Motivationen, Forschungsfrage, Methode, Vorgehen, Ergebnisse, Referenzen

Streaming Processing

- Betrugserkennung; wir wollen die eingehenden Adressen vergleichen mit der ursprünglichen Adresse der Pakete.
- Erkennung des besten Zustellers des Tages
- Analyse der Informationen über die Touren (Zustellerausgleich)
- Erreichbare Kunden analysieren

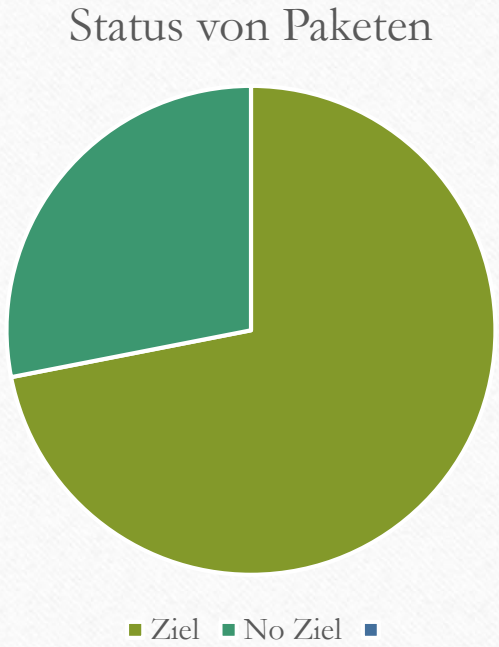
Was ist Big Data, Thema, Bisherige Arbeit, Motivationen, Forschungsfrage, Methode, Vorgehen, Ergebnisse, Referenzen

Consumer-API für Daten-Verbrauchen festlegen

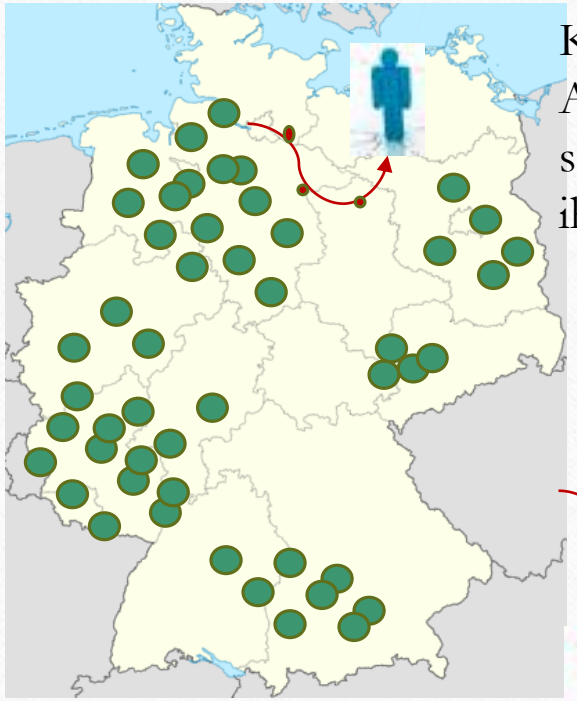
- Als Consumer kann man weitere Apps haben.

Mögliche Visulisierung
(**Positive Konsequenzen der Einführung von Streaming**)

Mehr Transparenz bei der Sendung von Paketen (Ziel)



Kunde bekommt genug Info und dadurch ist die Akzeptanz der Verspätung hoch



Kunde sieht die Annäherung von seinem Paket zu ihm.

Geo-Visualisierung von Paketen .

Strecke

Kunde

Was ist Big Data, Thema, Bisherige Arbeit, Motivationen, Forschungsfrage, Methode, Vorgehen, Ergebnisse, Referenzen

Ergebnisse (Mehrwerte unseres Konzepts aus den gegebenen Ergebnissen ableiten)

- Live Datenauswertung
- Live Optimierung (z.B. bei Geo-Visualisierung)
- Schnelle Entscheidungen (Kunde sieht die Fahrerposition und kann entgegen kommen, sich bei dem Fahrer authentifizieren und das Paket annehmen)
- Schnelle Reaktionszeit bei Fehlern in der Geschäftslogik (auf der Seite des Unternehmens)

Referenzen

- [1] Abdelkarim Ben Ayed, u.a , “Big Data Analytics for Logistics and Transportation ”, International Conference on Advanced logistics and Transport (IEEE’ ICAl T 2015)
- [2] https://blogs.informatica.com/2017/04/05/big-data-moving-from-technology-to-business-value-delivery/#fbid=FYtM6-w846L
- [3] Yi-Hsim, u.a , “Streaming Analytics Processing in Manufacturing Performance Monitoring and Prediction ”, International Conference on Big Data (IEEE’ BIGDATA 2017)
- [4] V. A. Ames, J. Gililland, J. Konopka, R. Schnabl, and K. Barber, Semiconductor Manufacturing Productivity Overall Equipment Effectiveness (OEE) Guidebook, Revision 1, Technology Transfer # 95032745A-GEN, SEMATECH, April 1995.
- [5] M. O’Neill, and P. Young, Developing Overall Equipment Effectiveness Metrics for Prototype Precision Manufacturing, PhD Thesis, Dublin City University, January 2011.
- [6] Manufacturing Analytics: Uncovering Secrets on Your Factory Floor, Sight Machine White Paper, Sight Machine, 2015.

Referenzen

- [7] https://kafka.apache.org/documentation/
- [8] Paul Le Noac'h u.a, "A Performance Evaluation of Apache Kafka in Support of Big Data Streaming Applications ", International Conference of Big Data (IEEE' BIGDATA 2017)
- [9]https://aws.amazon.com/de/streaming-data/
- Ovidiu-Cristian Marcu u.a, "Spark versus Flink: Understanding Performance in Big Data Analytics Frameworks ", International Conference on Cluster Computing (IEEE' ICCC 2016)
- Martin Andreoni Lopez u.a, "A Performance Comparison of Open-Source Stream Processing Platforms ", (IEEE' 2016)
- Bilal Akil u.a, "On the Usability of Hadoop MapReduce, Apache Spark & Apache Flink for Data Science ", International Conference of Big Data (IEEE' BIGDATA 2017)
- Paul Le Noac'h u.a, "A Performance Evaluation of Apache Kafka in Support of Big Data Streaming Applications ", International Conference of Big Data (IEEE' BIGDATA 2017)
- Wilhelmus Andrian Tanujaya u.a, "Rapid Data Stream Application Development Framework ", International Conference on Data and Software Engineering (IEEE' ICoDSE 2017)

Was ist Big Data, Thema, Bisherige Arbeit, Motivationen, Forschungsfrage, Methode, Vorgehen, Ergebnisse, Referenzen

Danke für eure Aufmerksamkeit