

Stream Processing

Life doesn't happen in batches...

Agenda

1. Einleitung

- a. Daten Merkmale
- b. Konzept “Bounded vs. Unbounded Data”

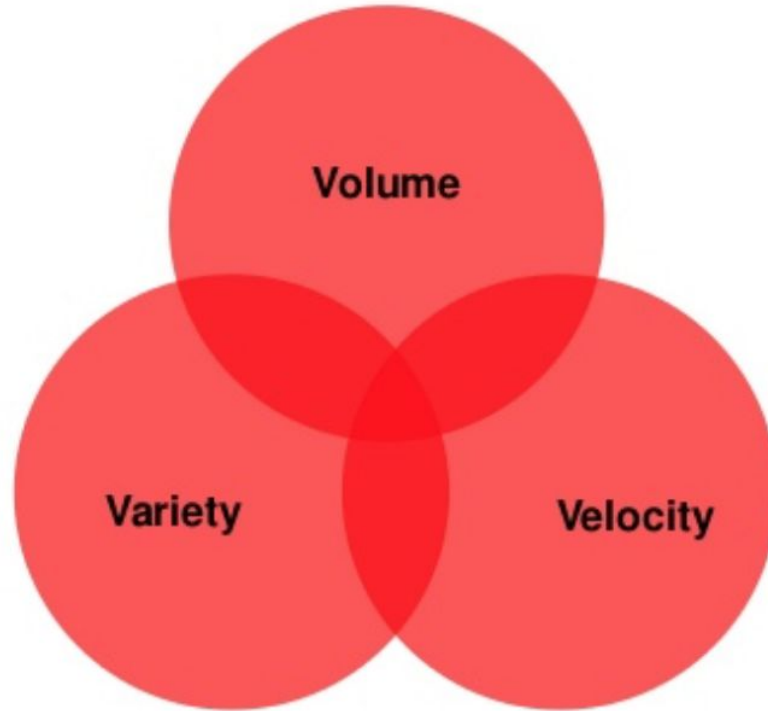
2. Batch- und Stream- Processing

- a. Architektur Muster
- b. Beispiel Google Analytics

3. Wann Stream Processing

4. Konzepte der Zeit in unbounded Data

1.a Daten Merkmale



1.b bounded vs. unbounded Data

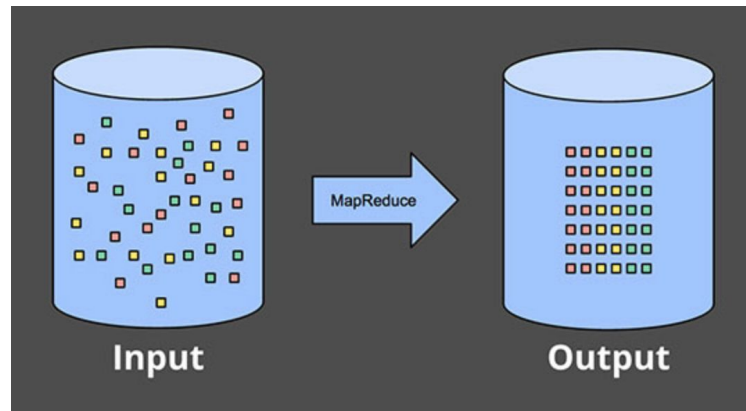
- **bounded data**
 - Datensatz ist finit (Bsp.: Cifar-10)
 - auf einem Medium persistiert
 - 'data-at-rest' genannt
- **unbounded data**
 - Datensatz ist infinit (Bsp.: Sensor-Daten)
 - kontinuierlicher Eingang von Daten (Streams)
 - 'data-in-motion' genannt

2. Batch- und Stream- Processing

2.a Architektur Muster

Batch-Processing

- Batch (Stapel, Bündel)
- Input: Bounded Data
- Output: Bounded Data
- Transformationen und Aktionen
- Verteilte Verarbeitung durch Framework

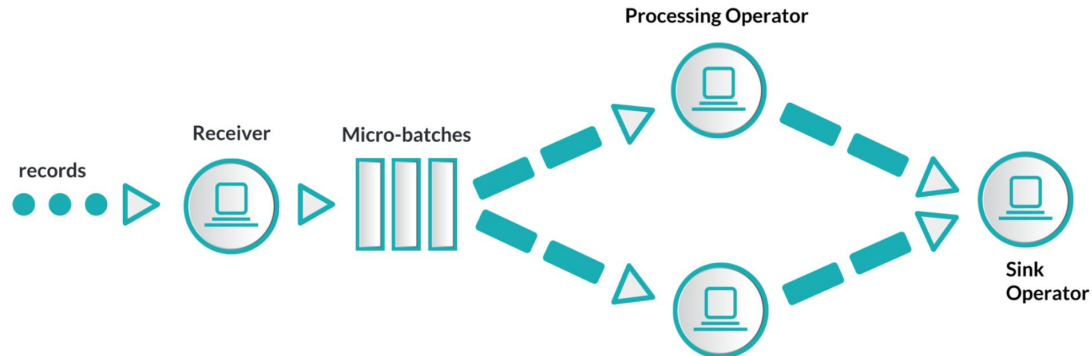


Bsp: `rdd.filter(_).map(x->x*2).sum()`

2.a Architektur Muster

Stream-Processing

- Stream (Strom, fließend)
- Input: Unbounded Data
- Output: Unbounded Data
- Verteilung der Last über Zeit
- Verteilte Verarbeitung innerhalb Operatoren



2.b Beispiel

- Google Analytics
 - Wie oft wird die Seite X aufgerufen
 - von IP-Adresse Y
 - mit Browser Z
 - im Zeitraum von/bis
 - ...

```
{  
  eventType: PageViewEvent,  
  timestamp: 1413215518,  
  ipAddress: 12.34.56.78,  
  sessionId: 106d2a521d3c6abcf36,  
  pageUrl: /talks.html,  
  referrer: google.com/search?q=...,  
  browser: Chrome 39  
}
```

2.b Batch-Processing Ansatz

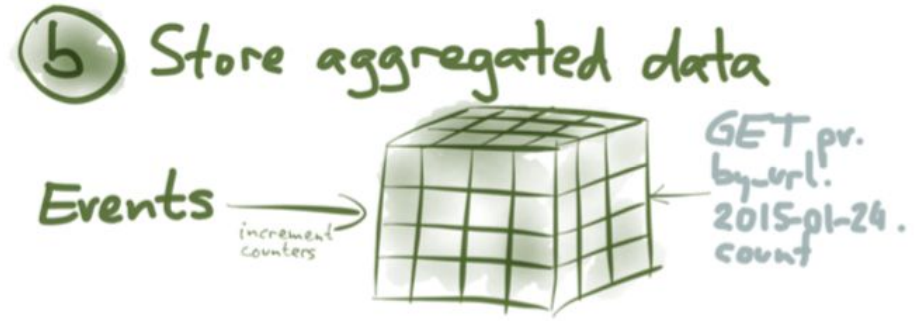
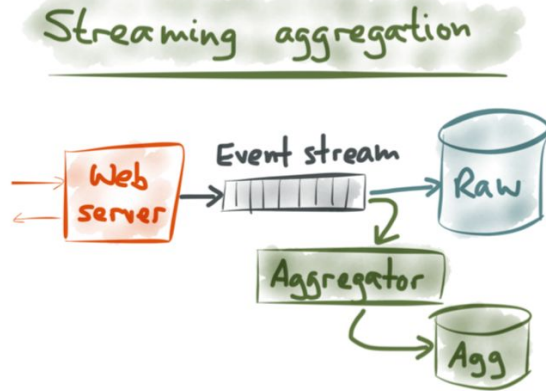
- Batch = DWH zum Zeitpunkt der Anfrage
- Processing = `Select Count(a) From Events Where .....`

```
context = HiveContext(sc)
results = context.sql(
    "SELECT * FROM people")
names = results.map(lambda p: p.name)
```



2.b Stream-Processing Ansatz

- Stream = Events entstehen kontinuierlich
- Processing = Zähler Inkrementieren



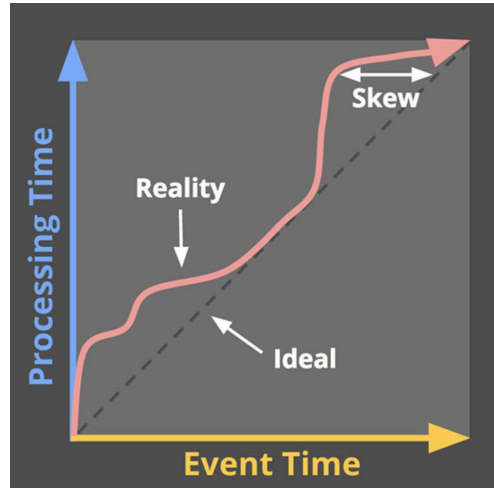
3. Wann Stream Processing

- Velocity
 - “event happens” -> “insight derived” -> “action is taken”
 - geringe Latenz zwischen Ereignis und Ergebnis Notwendig
- Volumen
 - Streams verarbeiten/analysieren Daten direkt
 - “Bedeutungslose Daten” filtern
 - “Bedeutungsvolle Daten” weiterleiten und/oder persistieren
- Variety
 - mehrere Streams
 - unterschiedliche Datenquellen
 - unterschiedliche Datenstrukturen

4. Konzepte der Zeit in “Unbounded Data”

4. Event time vs. Processing time

- Event time
 - Zeitpunkt an dem das Event tatsächlich aufgetreten ist.
- Processing time
 - Zeitpunkt an dem das Event im System beobachtet wurde..



weiterführende Quellen

The world beyond batch: Streaming 101 und 102:

<https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101#section-ref>

<https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-102>

Weitere Referenzen:

<http://www.oreilly.com/data/free/stream-processing.csp>

<https://www.cakesolutions.net/teamblogs/comparison-of-apache-stream-processing-frameworks-part-1>

<https://vsis-www.informatik.uni-hamburg.de/getDoc.php/publications/561/Real-time%20stream%20processing%20for%20Big%20Data.pdf>

<https://blog.codecentric.de/2017/03/verteilte-stream-processing-frameworks-fuer-fast-data-big-data-ein-meer-moeglichkeiten/>