Big-Data-Technologien

Kapitel 2: Batch-Verarbeitung

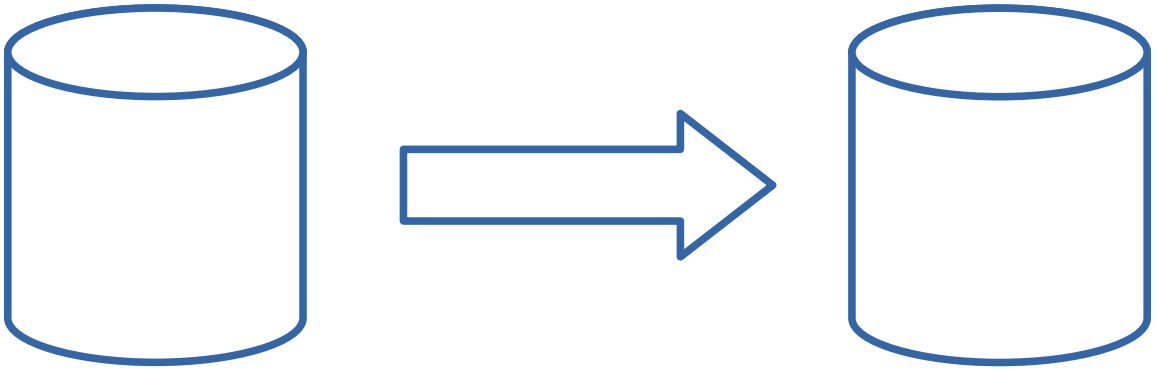
Hochschule Trier
Prof. Dr. Christoph Schmitz

Überblick

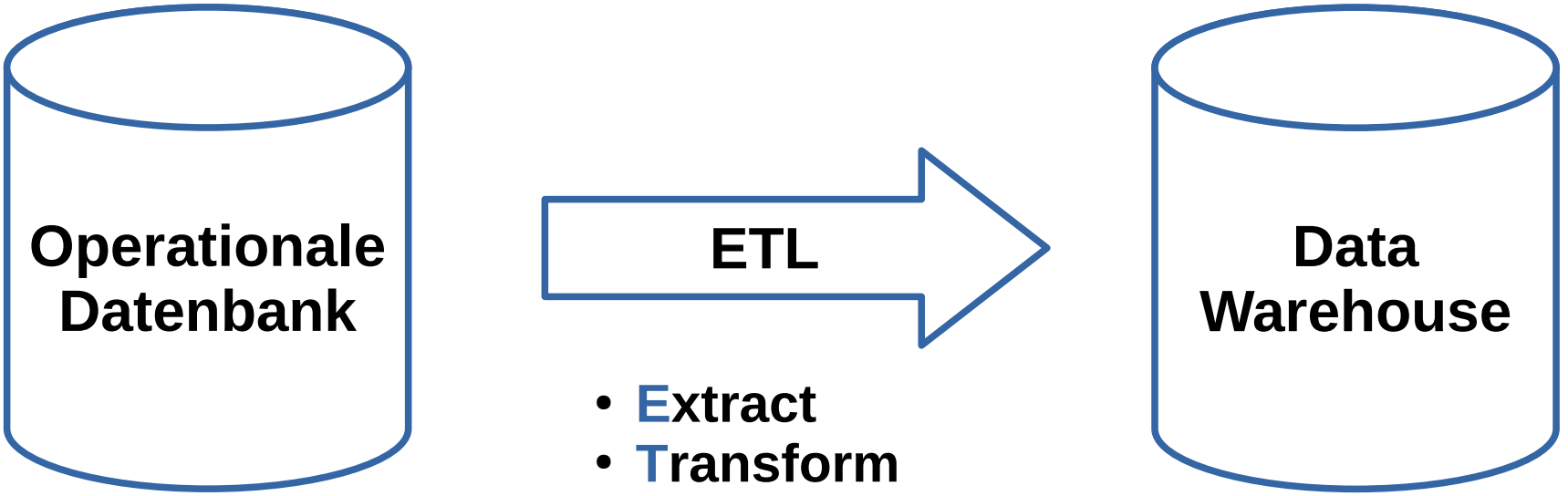
- Batch-Verarbeitung
- Einführung Apache Hadoop

Batch-Verarbeitung

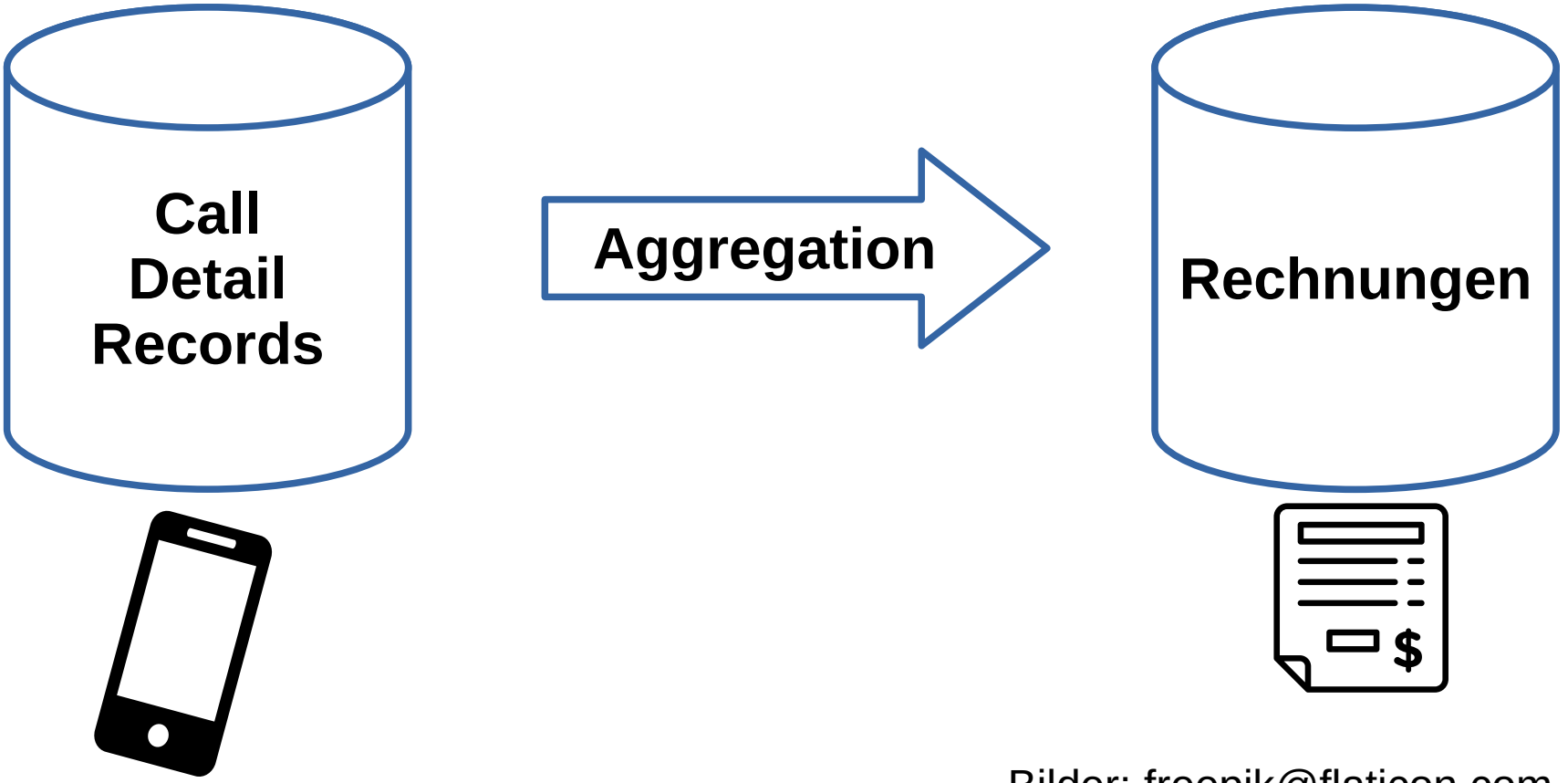
- Verarbeiten sehr großer Datenmengen als Ganzes
- Möglichst hoher Durchsatz
- Latenz von Minuten/Stunden



Beispiel: Batch-Verarbeitung

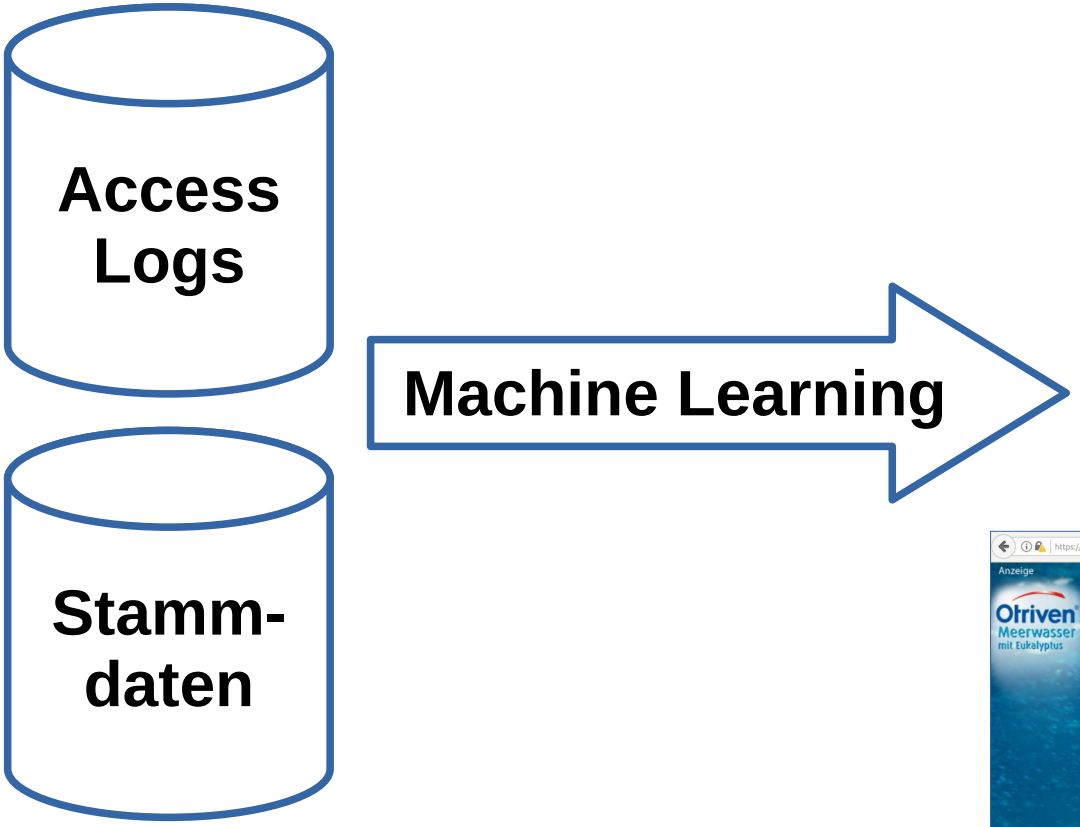


Beispiel: Batch-Verarbeitung



Bilder: freepik@flaticon.com

Beispiel: Batch-Verarbeitung



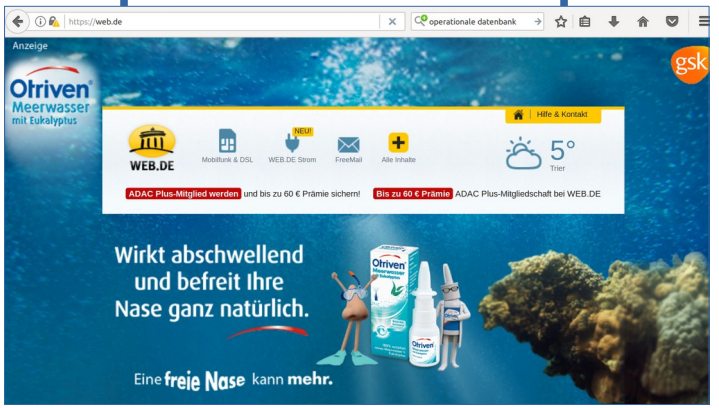


Bild: web.de

Apache Hadoop

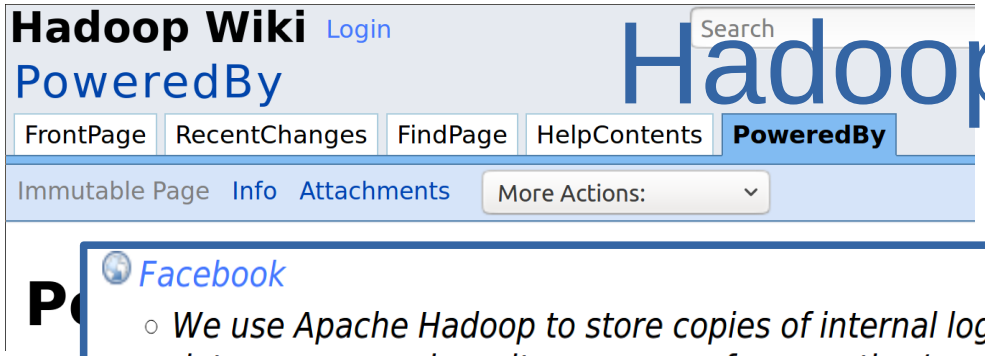
Apache Hadoop

- Hadoop ist...
  - ein **Application Framework** für verteilte Datenverarbeitung
  - Big-Data-**Infrastruktur**
  - ein Big-Data-**Ökosystem**

Hadoop: Geschichte

- Entstanden aus **Nutch**, ca. 2005
- Basiert auf Ideen von Google
  - MapReduce (2004)
  - Google File System (2003)
- Weiterentwicklung z. B. bei Yahoo
- Heute in Form von **Distributionen**
  - **Cloudera**
  - **Hortonworks**
  - **MapR**
  - Amazon AWS, IBM, Microsoft Azure, ...

Hadoop: Zahlen



P

Facebook

- *We use Apache Hadoop to store copies of internal log and dimension data sources and use it as a source for reporting/analytics and machine learning.*
- *Currently we have 2 major clusters:*
  - *A 1100-machine cluster with 8800 cores and about 12 PB raw storage.*
  - *A 300-machine cluster with 2400 cores and about 3 PB raw storage.*
  - *Each (commodity) node has 8 cores a*

Yahoo!

- *More than 100,000 CPUs in >40,000 computers running Hadoop*
- *Our biggest cluster: 4500 nodes (2*4cpu boxes w 4*1TB disk & 16GB RAM)*
  - *Used to support research for Ad Systems and Web Search*
  - *Also used to do scaling tests to support development of Apache Hadoop on larger clusters*

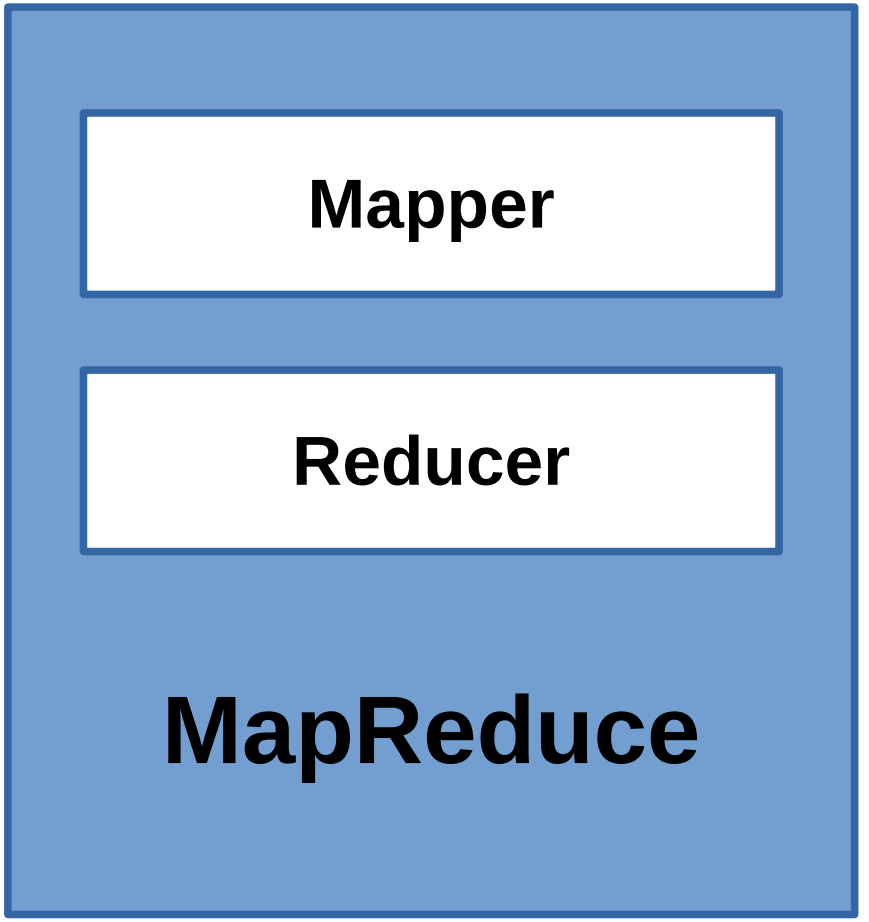
Spotify

- We use Apache Hadoop for content ger
  aggregation, reporting, analysis (see m
  Evolution of Hadoop at Spotify - Throug
  Pain) and even for generating music re
  How Apache Drives Music Recommendations At Spotify
- 1650 node cluster : 43,000 virtualized cores, ~70TB RAM, ~65 PB storage (read more about our Hadoop issues while growing fast: Hadoop Adventures At Spotify)
- +20,000 daily Hadoop jobs (scheduled by Luigi, our open-sourced job orchestrator - code and video)

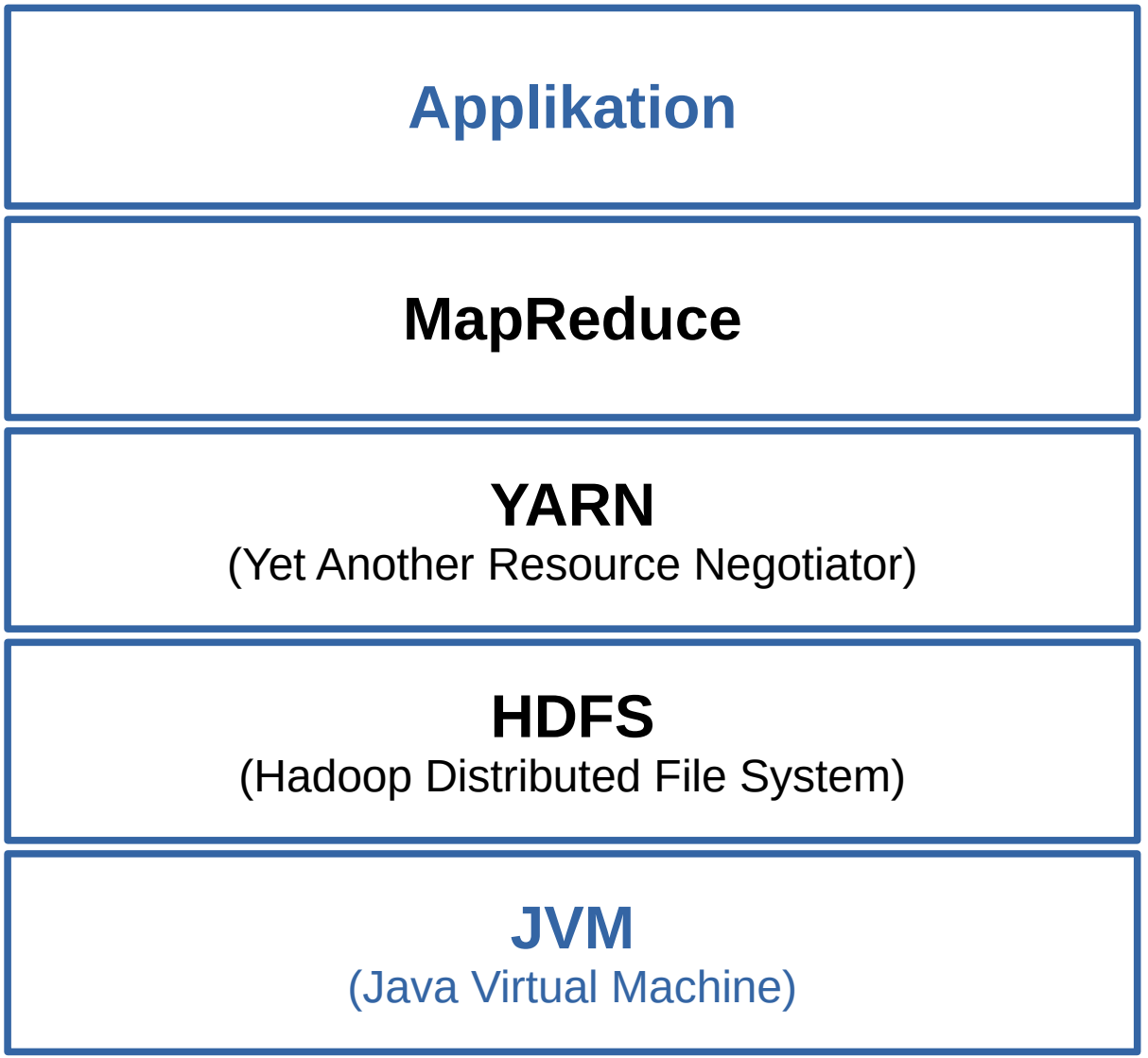
Hadoop als Application Framework

MapReduce bietet:

- Progammiermodell
- Automatische Verteilung
- Fehlertoleranz
- Monitoring



Hadoop als Infrastruktur

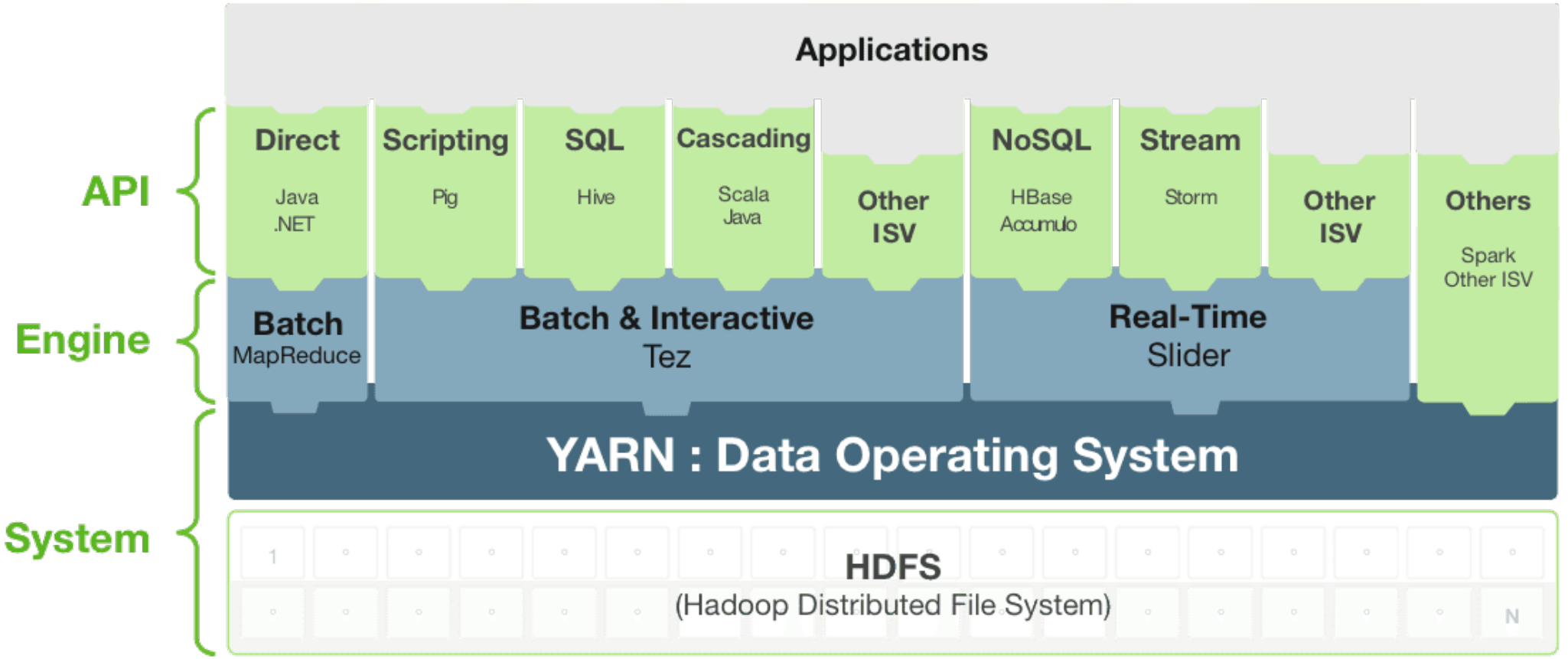


Programmiermodell
für verteiltes Rechnen

Ressourcenverwaltung
(CPUs, Speicher)

Verteilte, redundante
Datenhaltung

Hadoop als Ökosystem



Quelle: Hortonworks