Real world data and Big Data Analytics Platform

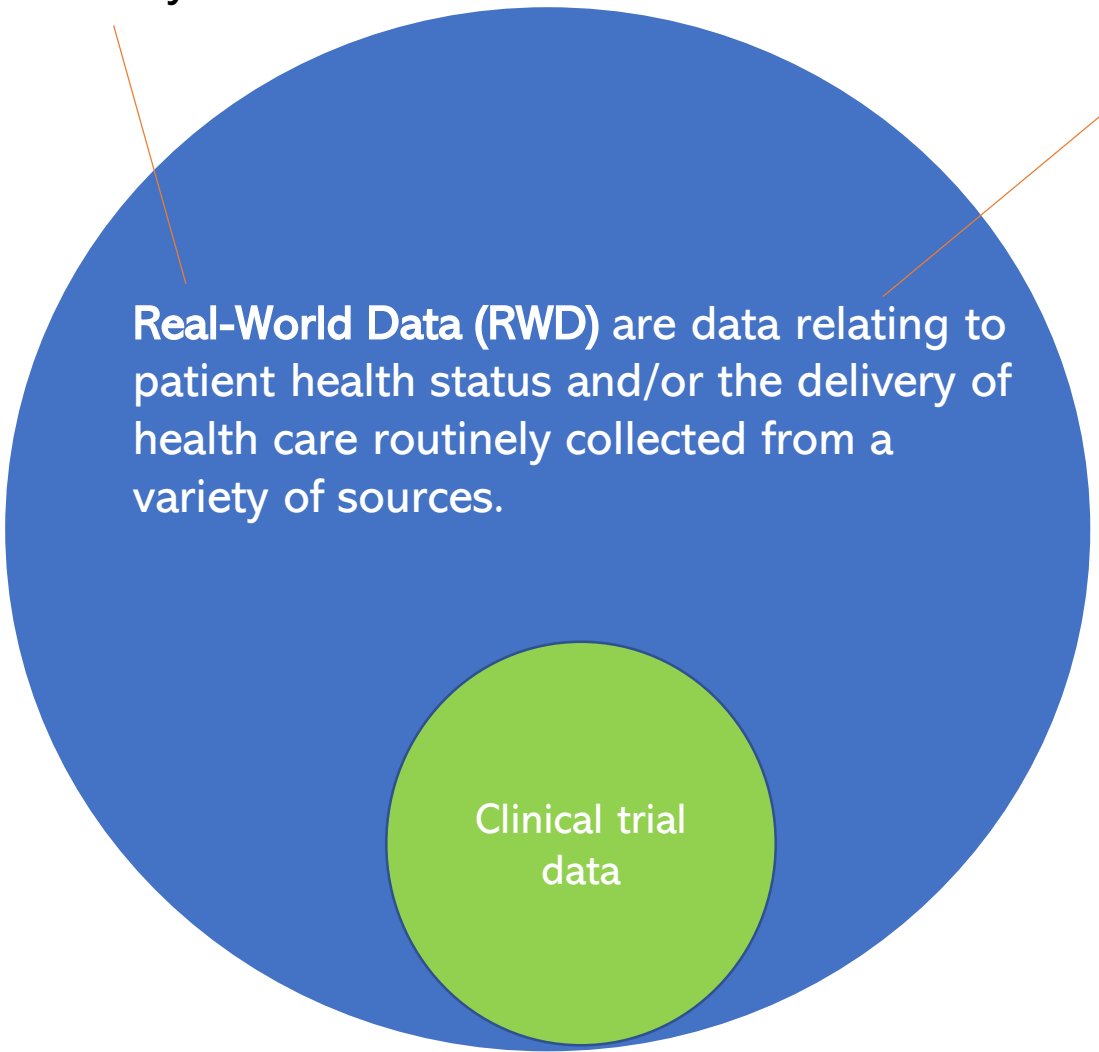
Current guidelines on medical device RWD

- FDA
  - 2017, Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices
  - 2018, Framework for FDA's Real-World Evidence Program
  - 2018, Use of Electronic Health Record Data in Clinical Investigations Guidance for Industry

- NMPA
  - 2020, 真实世界数据用于医疗器械临床评价技术指导原则

Real world data

RWE: clinical evidence regarding the usage and analysis of RWD



RWD (structured, semi-structured, unstructured)

- Electronic health records (EHRs)
- Claims and billing activities
- Product and disease registries
- Patient-generated data including in home-use settings
- Data gathered from other sources that can inform on health status, such as mobile devices

Key uses of RWE for medical device

- Expanded indications for use
- Post-market Surveillance Studies
- Objective performance criteria and performance goals
- Comparative effectiveness research (comparison of multiple devices)

Big Data for Big Pharma

	Big Data for R&D	Therapeutic/ Research Areas	Alliances/ Partners	Clinical trial data transparency	Data technology infrastructure	Best Practice Case Study
NOVARTIS	✓	• Oncology • Ophthalmology • Haematology (Multiple Sclerosis) • NGS	• Google • Covance	✓	• MapR • HTS Explorer • Chemotopgraphy • ConTour	• Detection of glomerulo-sclerosis as cause of kidney cancer
Pfizer	✓	• Oncology • Fibromyalgia • Obesity • Biomarker-focused research • Personalized medicine • NGS	• Humedica • CliniWorks • Optum Labs	✗	• Precision Medicine Analytics Ecosystem	• Xalkori
SANOFI	✓	• NGS • Translational medicine	• NextBio • IBM Watson	✓	—	✗
Roche	✓	• Oncology • NGS • RWD/RWE	• Bina Technologies • Foundation Medicine • Astra Zeneca • Point Cross	✓	• Cloudera Impala • Hadoop	✗
MERCK	✓	• Oncology • Vaccines • Animal Health	• Smart Patients • Practice Fusion • Allscripts • DGI	✗	• GsDesign Explorer • Hadoop	• Optimization of manufacturing of vaccines

Big Data Analytics Company

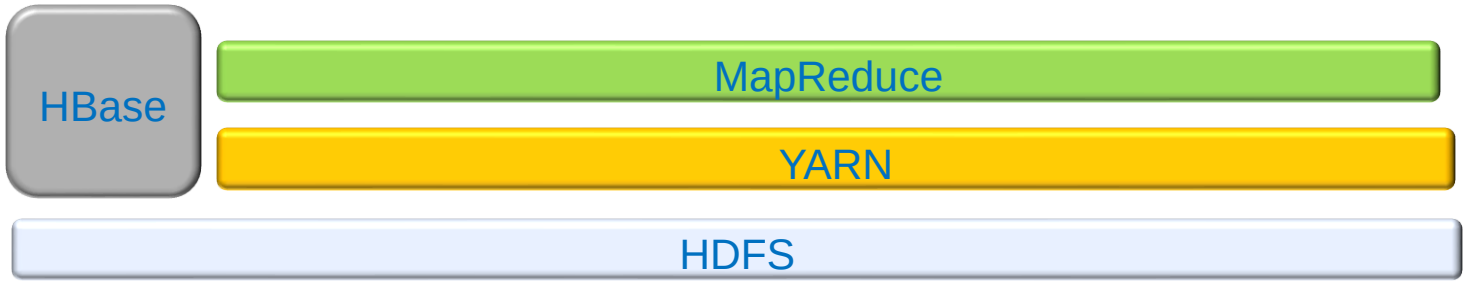
Cloudera
Hortonworks
MapR
MongoDB
Amazon
Google

Big Data Analytics Company	Pharmaceutical company
MapR	Novartis, Boehringer Ingelheim
Cloudera	Roche, IQVIA, Celgene, GSK
Hortonworks	Merck
MongoDB	AstraZeneca, Medtronic
Amazon	Amgen, Alcon, Merck

Hadoop

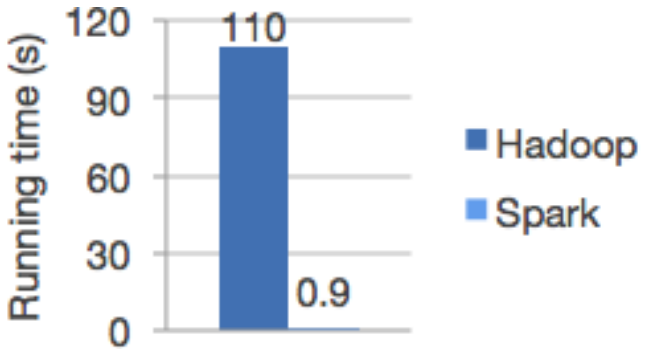
Hadoop is a software ecosystem that allows for massively parallel computing

- **Hadoop Distributed File System (HDFS)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **HBase (NoSQL)**



Spark

- Spark is a unified analytics engine for large-scale data processing



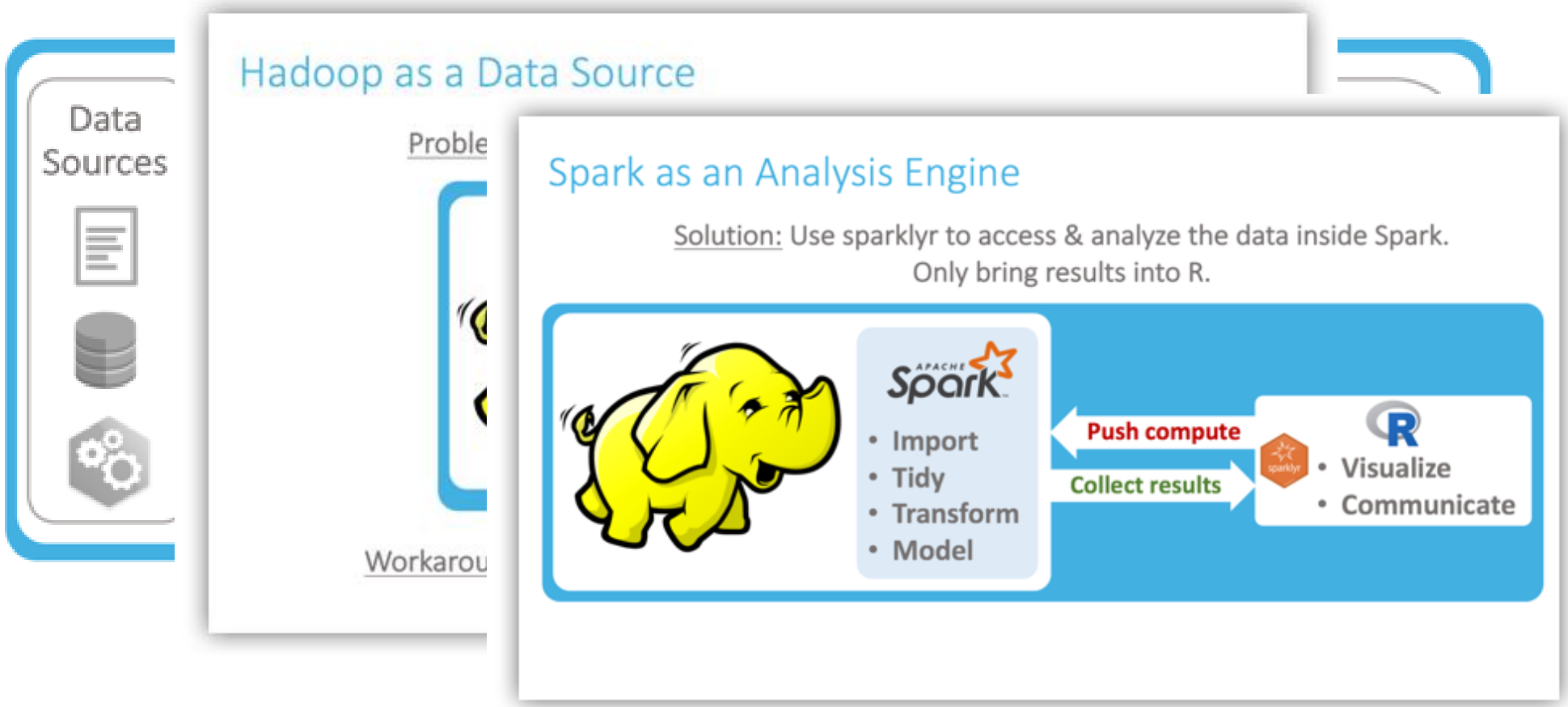
Key Features	Apache Spark	Hadoop MapReduce
Speed	10–100 times faster than MapReduce	Slower
Analytics	Supports streaming, Machine Learning, complex analytics, etc.	Comprises simple Map and Reduce tasks
Suitable for	Real-time streaming	Batch processing
Coding	Lesser lines of code	More lines of code
Processing Location	In-memory	Local disk

Data Analytic Tools

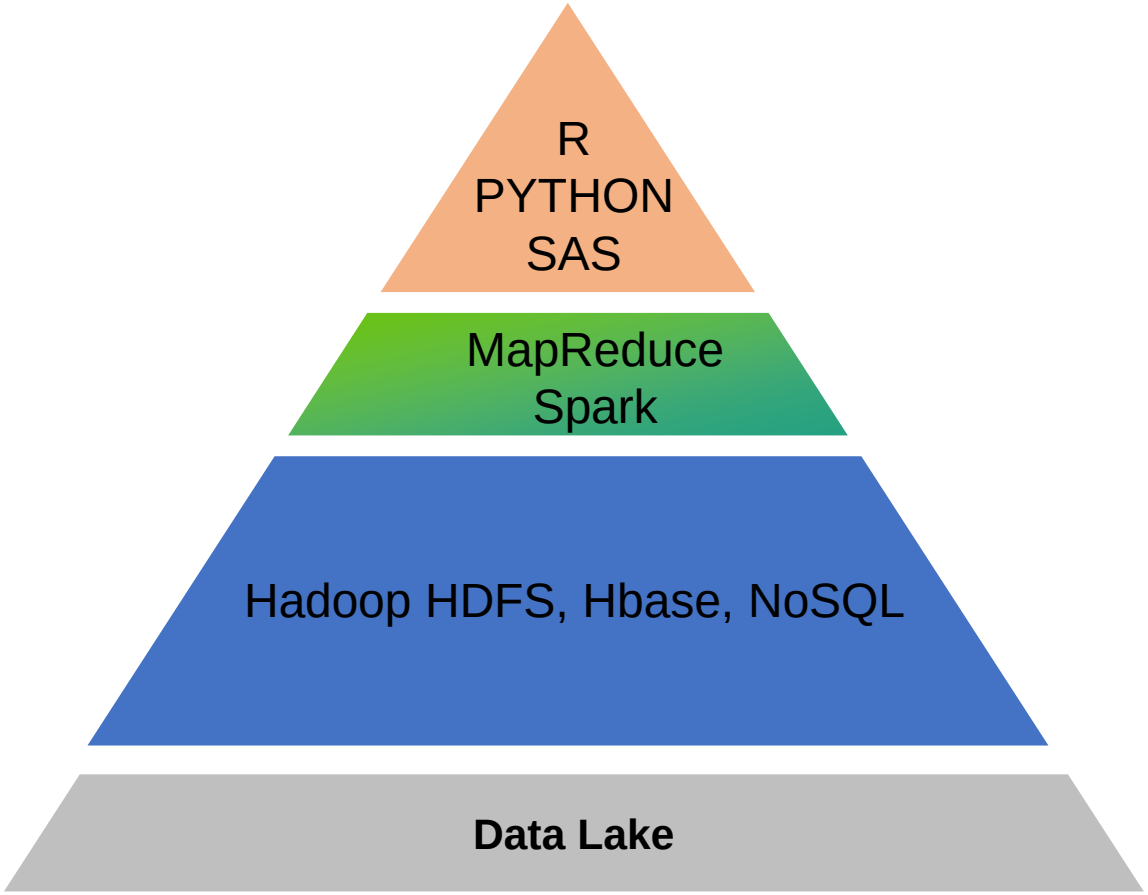
- SAS
- R
- Python
- TIBCO Spotfire
- Tableau

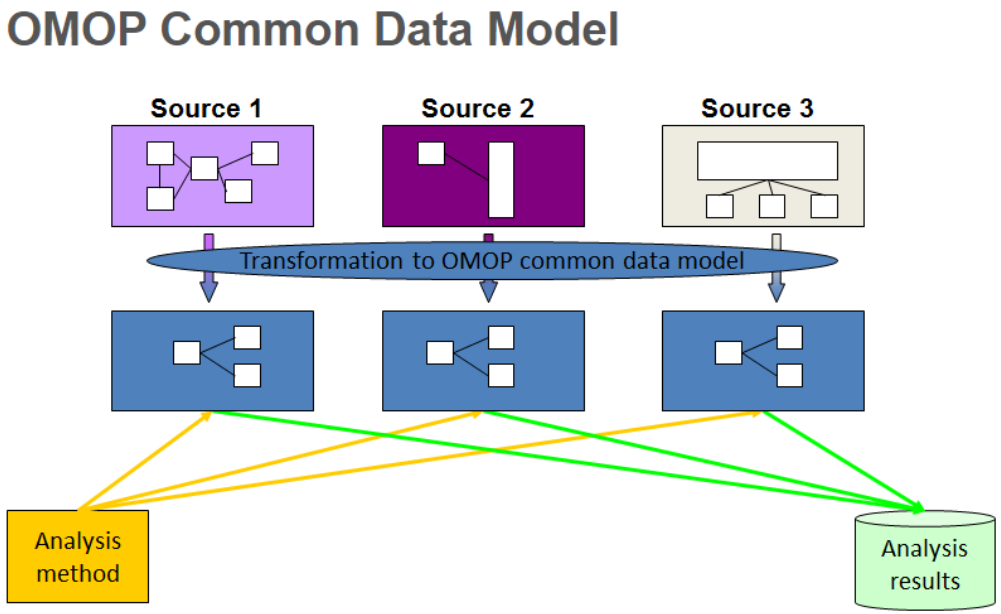
Big Data Analysis workflow

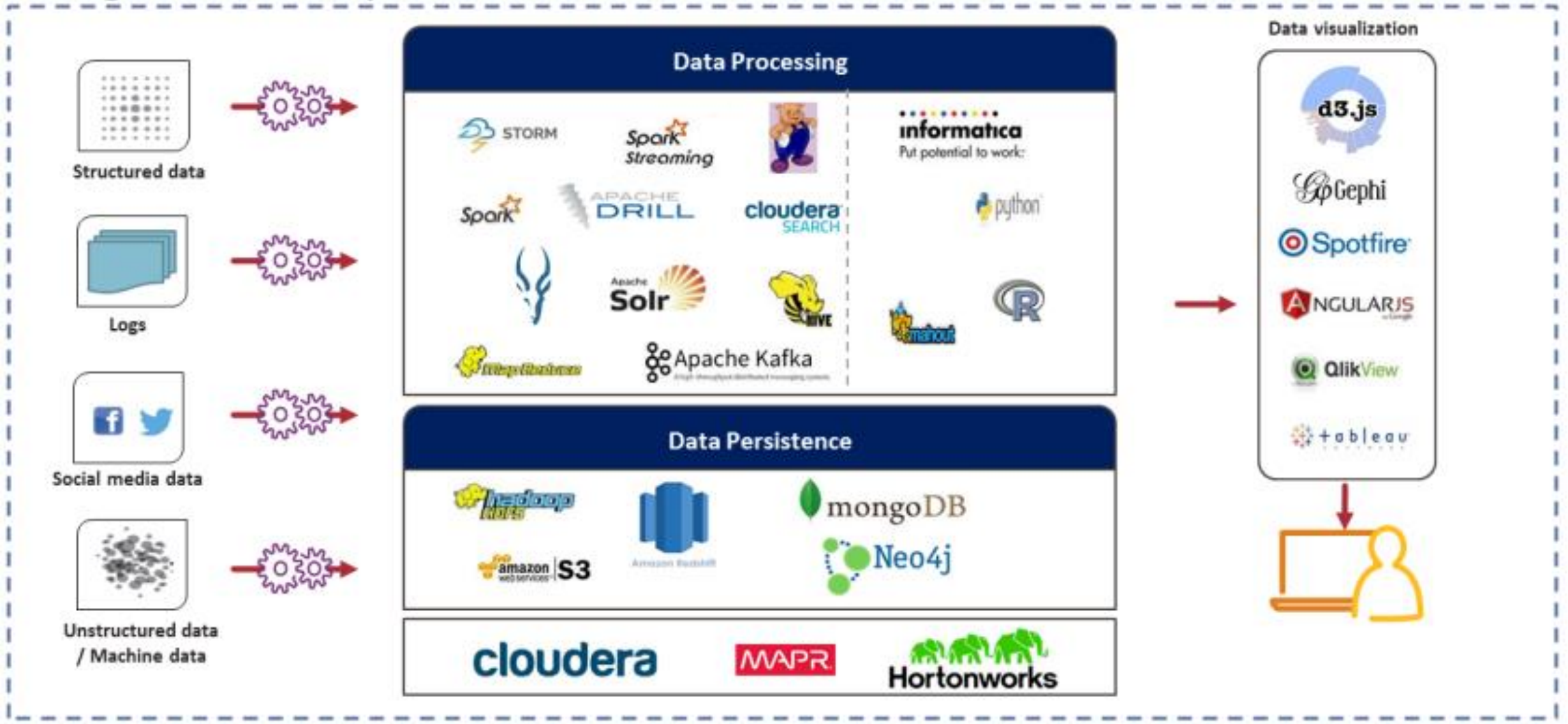
R for Data Science



Big Data Analytics Framework






Structured data
Logs
Social media data
Unstructured data / Machine data
Data Processing
STORM
Spark Streaming
Spark
APACHE DRILL
cloudera SEARCH
Apache Solr
Apache Kafka
informatica
Put potential to work:
python
mahout
Data Persistence
hadoop
amazon web services S3
mongoDB
Neo4j
cloudera
MAPR
Hortonworks
Data visualization
d3.js
Gephi
Spotfire
ANGULARJS
QlikView
tableau

Example of RWD sources to support MD RWE

ADMINISTRATIVE DATABASES

Examples*:

- **Publicly Available**
  - Healthcare Cost and Utilization Project (HCUP) (eg, Nationwide Inpatient Sample)
  - Medicare/Medicaid Standard Analytic Files
  - National Hospital Discharge Survey
  - Surveillance, Epidemiology, and End Results (SEER)—Medicare
- **Payer-sourced Data**
  - Optum
  - HealthCore/Anthem, Inc
  - Blue Health Intelligence
  - Korean Health Insurance Review and Assessment
- **Hospital/Group Purchasing Organization**
  - Premier Hospital Database
  - Vizient (formerly MedAssets) Database
  - MedMining/Geisinger
  - Japanese Medical Data Vision
- **Multisource Data Consolidations**
  - IBM Watson Health/Truven/MarketScan
  - IQVIA Pharmetrics
  - Japanese Medical Data Center (Japan)
  - Orizon (Brazil)

Key Considerations:

- Relatively inexpensive and rich in data elements like diagnoses, procedures, medications, and healthcare costs/ expenditures
- Typically comprise data from millions of patients and therefore are considered to have good generalizability
- Medical device identification is often dependent on the device possessing a specific billing code (eg, a Healthcare Common Procedure Coding System code), or mining unstructured data fields, such as hospital charge master data or physician notes, which can introduce measurement error
- Cannot usually answer questions such as why a provider chose one therapeutic approach over another (eg, surgery versus medication)
- Can lack information on important device-specific outcomes, such as device failures

ELECTRONIC HEALTH RECORDS

Examples*:

- Hospitals/academic medical centers
- Community practice sites
- Flatiron Health Oncology
- Cerner Health Facts
- Optum/Humedica
- US Oncology
- Practice Fusion
- GE Healthcare Centricity
- Clinical Practice Research Datalink (UK)
- IBM Watson Health Explorys

Key Considerations:

- Limited longitudinal follow-up, sometimes unable to track patients across sites of care
- Typically have same medical device identification challenges as administrative databases
- With proper design, researchers may be able to evaluate "why" events happen during treatment or treatment decision rationales
- Can lack information on important device-specific outcomes, such as device failures

SURVEYS & REGISTRIES

Examples*:

- Society of Thoracic Surgeons (STS) National Database
- Vascular Quality Initiative
- Japan PCI (Japan)
- US Cath-PCI Registry
- National Cardiovascular Data Registry's Implantable Cardiac Device Registry
- National Joint Replacement Registry (Australia)
- Kaiser Permanente National Total Joint Replacement Registry
- National Joint Registry (GB, Wales, N-IRL)
- Canadian Joint Replacement Registry
- Kaiser Permanente National Implant Registries
- European Database for Medical Devices (anticipated launch in 2020)

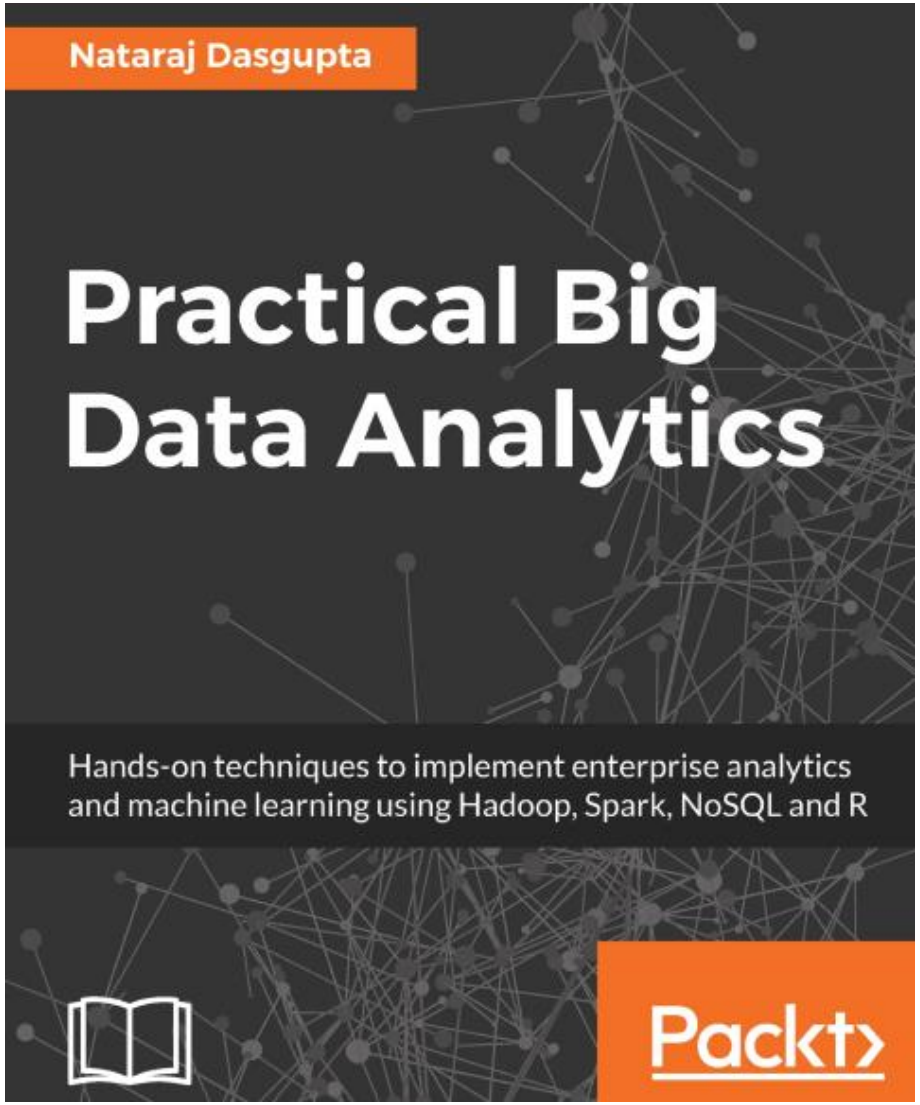
Key Considerations:

- Can collect and yield medical device satisfaction information directly from patients
- Provider surveys and expert panels can provide insights into clinical perspectives on drivers of treatment choice and product prescribing preferences
- Direct-to-subject study designs are often patient-centered and can capture subjective information unavailable via claims data or medical records
- Limited longitudinal follow-up; ability to link to other longitudinal data sources is inconsistent
- Information specific to the purpose of the registry design or to the remit of the expert panel is included, but they are otherwise limited in scope

References

- https://rweframework.com/
- https://spark.rstudio.com/guides/data-lakes/
- https://mapr.com/customers/
- http://www.big-datapharma.com/
- https://www.cloudera.com/about/customers/glaxosmithkline.html
- https://www.mongodb.com/blog/post/how-medtronic-manages-machine-data-mongodb?c=d67bc669bb
- https://medium.com/better-programming/hadoop-vs-hdfs-vs-hbase-vs-hive-ddfffd45d222
- https://www.lexjansen.com/phuse/2017/rw/RW02.pdf
- https://books.google.co.jp/books?id=xY40DwAAQBAJ&pg=PA1977&lpg=PA1977&dq=big+data+for+big+pharma,+google+book&source=bl&ots=WRd14KVrvZ&sig=ACfU3U0NsBPmxXas7Q_HcZPur3slhibcOw&hl=en&sa=X&ved=2ahUKEwijpOG058fqAhXNGaYKHTQTBtUQ6AEwBHoECAoQAQ#v=onepage&q&f=false
- https://www.ibmbigdatahub.com/blog/how-build-all-purpose-big-data-engine-hadoop-and-spark
- https://aws.amazon.com/blogs/apn/how-implementing-a-real-world-evidence-platform-on-aws-drives-real-world-business-value/
- https://mapr.com/resources/novartis-relies-mapr-flexible-big-data-solutions-drug-discovery/assets/mapr_customerbrief_novartis_040315.pdf
- https://www.ncbi.nlm.nih.gov/books/NBK441694/#__NBK441694_dtls__
- https://www.intechopen.com/books/big-data-on-real-world-applications/pesscara-an-example-infrastructure-for-big-data-research

Book


Nataraj Dasgupta
Practical Big Data Analytics
Hands-on techniques to implement enterprise analytics and machine learning using Hadoop, Spark, NoSQL and R
Packt


CHARITÉ
HUMBOLDT · UNIVERSITÄT
Schriftenreihe Masterstudiengang Consumer Health Care
herausgegeben von Prof. Dr. Marion Schaefer
Malena Johannes
Big Data for Big Pharma
An Accelerator for
The Research and Development Engine?
ibidem