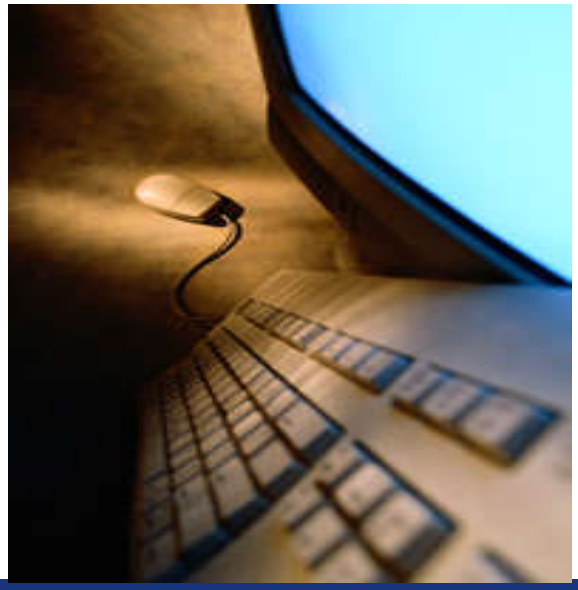
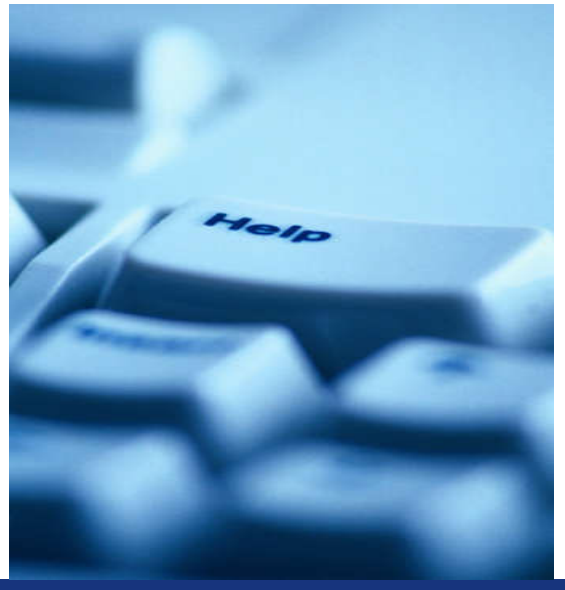
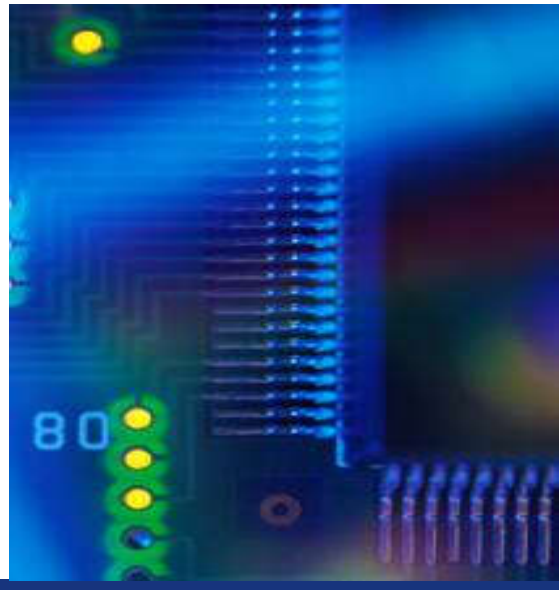
UNIVERSITAS
BUDI LUHUR







FAKULTAS TEKNOLOGI INFORMASI

REKAYASA DATA

Data engineering

- Data engineering is the development, operation, and maintenance of data infrastructure, either on-premises or in the cloud (or hybrid or multi-cloud), comprising databases and pipelines to extract, transform, and load data.

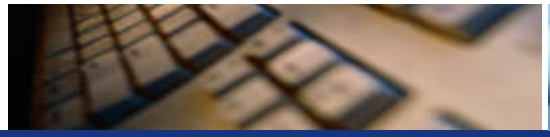
Data engineering versus data science

- Data engineering is what makes data science possible.
- Data scientists may be expected to clean and move the data required for analysis.
- Data scientists and data engineers use similar tools (Python, for instance), but they specialize in different areas. Data engineers need to understand data formats, models, and structures to efficiently transport data, whereas data scientists utilize them for building statistical models and mathematical computation.


Data engineering versus data science

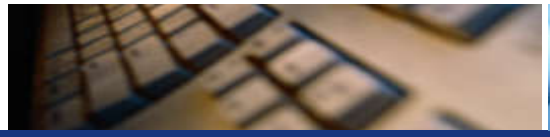
- Data scientists will connect to the data warehouses built by data engineers. From there, they can extract the data required for machine learning models and analysis.

- Data scientists may have their models incorporated into a data engineering pipeline. A close relationship should exist between data engineers and data scientists. Understanding what data scientists need in the data will only serve to help the data engineers deliver a better product.




What Data Engineers do ?

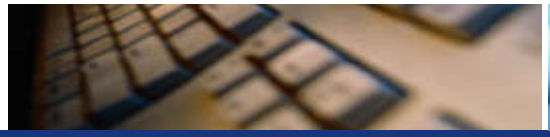
- Data engineering is part of the big data ecosystem and is closely linked to data science.
- Data engineers work in the background and do not get the same level of attention as data scientists, but they are critical to the process of data science.
- The roles and responsibilities of a data engineer vary depending on an organization's level of data maturity and staffing levels; however, there are some tasks, such as the extracting, loading, and transforming of data, that are foundational to the role of a data engineer.




What Data Engineers do ?

- An online retailer has a website where you can purchase widgets in a variety of colors. The website is backed by a relational database. Every transaction is stored in the database. How many blue widgets did the retailer sell in the last quarter?

- To answer this question, you could run a SQL query on the database. This doesn't rise to the level of needing a data engineer.



What Data Engineers do ?

- But as the site grows, running queries on the production database is no longer practical.
- Furthermore, there may be more than one database that records transactions.
- There may be a database at different geographical locations – for example, the retailers in North America may have a different database than the retailers in Asia, Africa, and Europe.

- Now you have entered the realm of data engineering.

What Data Engineers do ?

- **To answer the preceding question, a data engineer would create connections to all of the transactional databases for each region, extract the data, and load it into a data warehouse. From there, you could now count the number of all the blue widgets sold.**

- **Rather than finding the number of blue widgets sold, companies would prefer to find the answer to the following questions:**
  - How do we find out which locations sell the most widgets?
  - How do we find out the peak times for selling widgets?
  - How many users put widgets in their carts and remove them later?
  - How do we find out the combinations of widgets that are sold together?


What Data Engineers do ?

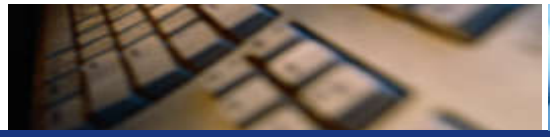
- Answering these questions requires more than just extracting the data and loading it into a single system.

- There is a transformation required in between the extract and load.
- There is also the difference in times zones in different regions. For instance, the United States alone has four time zones. Because of this, you would need to transform time fields to a standard.
- You will also need a way to distinguish sales in each region. This could be accomplished by adding a location field to the data. Should this field be spatial – in coordinates or as well-known text – or will it just be text that could be transformed in a data engineering pipeline?

What Data Engineers do ?

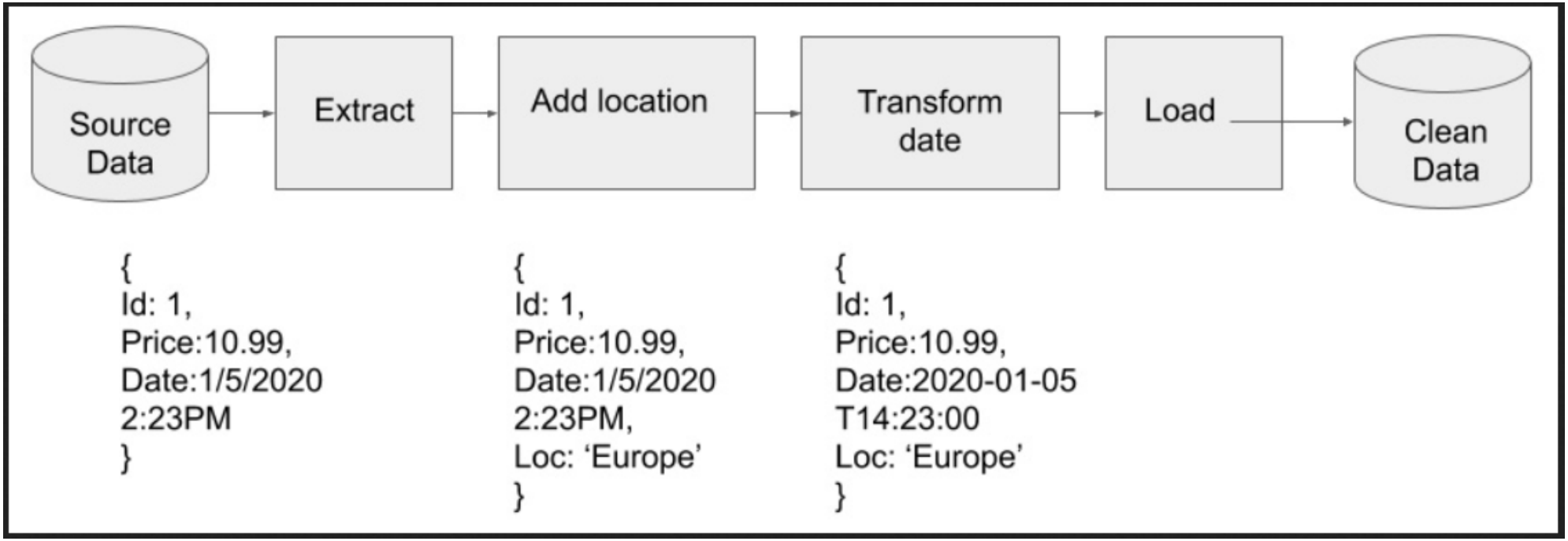
- **Here, the data engineer would need to extract the data from each database, then transform the data by adding an additional field for the location. To compare the time zones, the data engineer would need to be familiar with data standards. For the time, the International Organization for Standardization (ISO) has a standard – ISO 8601.**

- **Let's now answer the questions in the preceding list one by one:**
  - Extract the data from each database.
  - Add a field to tag the location for each transaction in the data
  - Transform the date from local time to ISO 8601.
  - Load the data into the data warehouse.

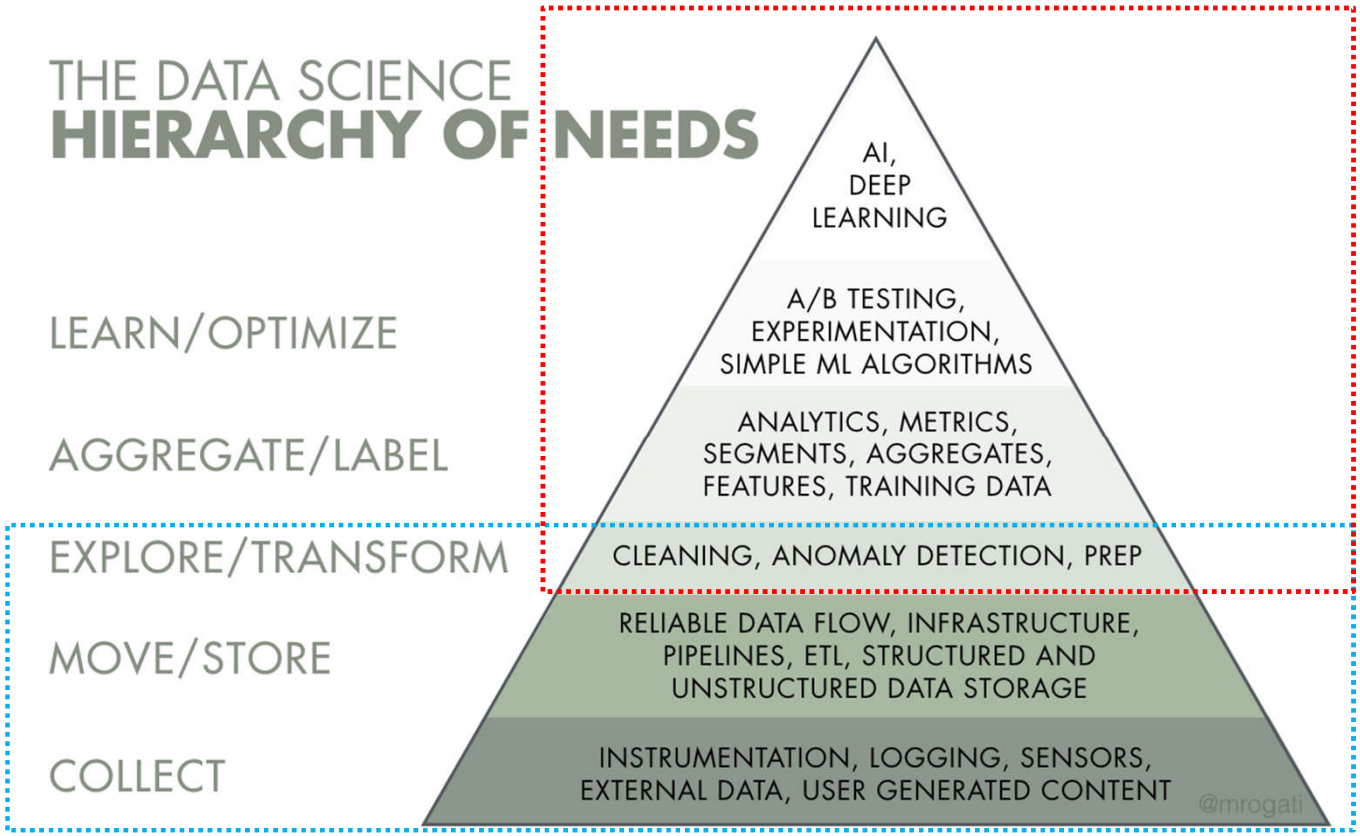



What Data Engineers do ?

- The combination of extracting, loading, and transforming data is accomplished by the creation of a data pipeline.
- The data comes into the pipeline raw, or dirty in the sense that there may be missing data or typos in the data, which is then cleaned as it flows through the pipe.
- After that, it comes out the other side into a data warehouse, where it can be queried. The following diagram shows the pipeline required to accomplish the task:


Source Data
Extract
Add location
Transform date
Load
Clean Data
{
Id: 1,
Price:10.99,
Date:1/5/2020
2:23PM
}
{
Id: 1,
Price:10.99,
Date:1/5/2020
2:23PM,
Loc: 'Europe'
}
{
Id: 1,
Price:10.99,
Date:2020-01-05
T14:23:00
Loc: 'Europe'
}

Hierarchy of Needs



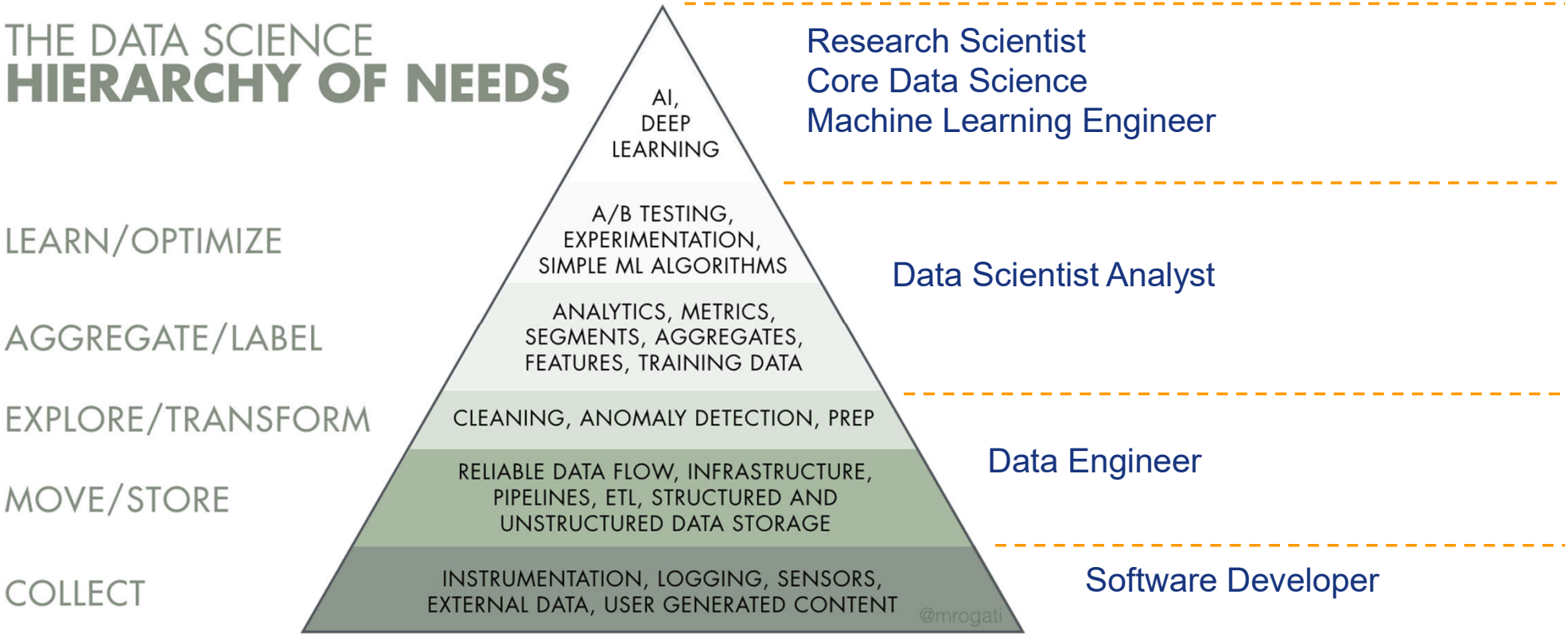
Data Scientist

Data Engineer

Sumber: https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Sumber: Insinyur Data

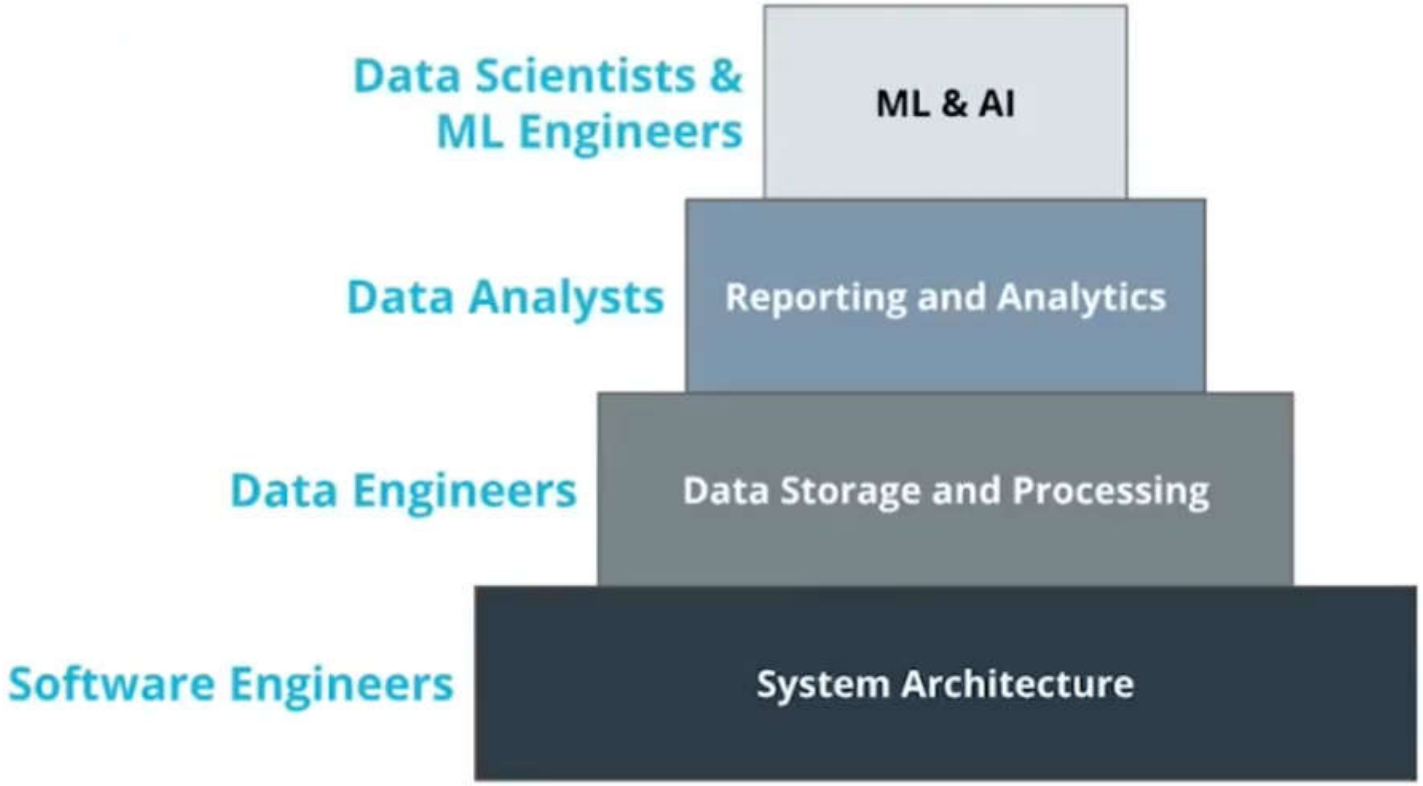
Hierarchy of Needs



Sumber: https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

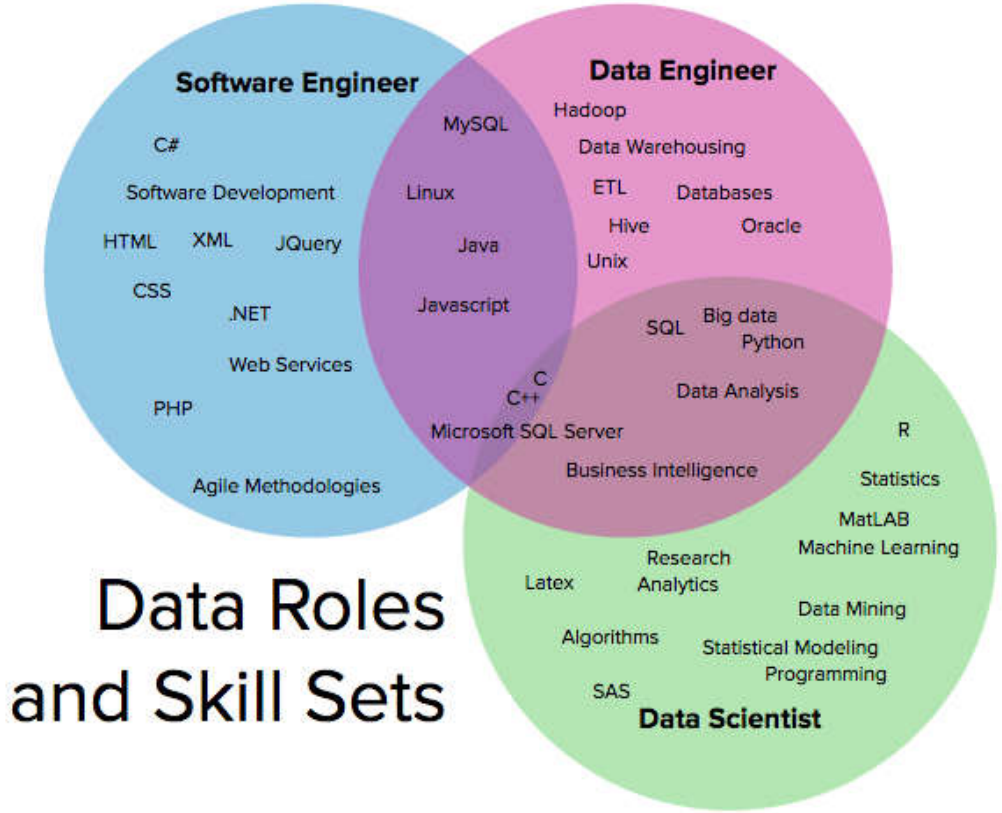
Sumber: Data Engineering Nanodegree v1.0.0 - Udacity

Data Roles and Skill Sets



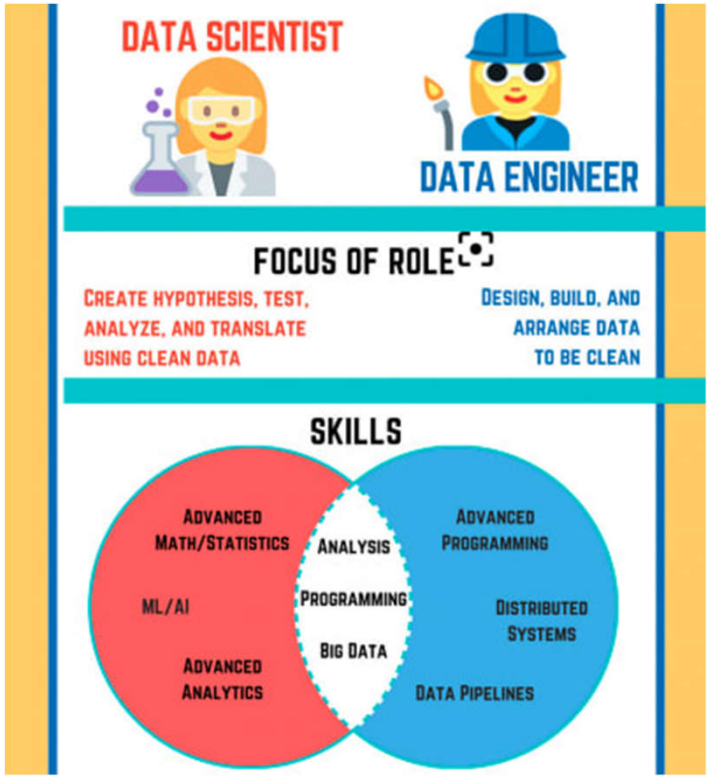
Sumber: Data Engineering Nanodegree v1.0.0 - Udacity

Data Roles and Skill Sets



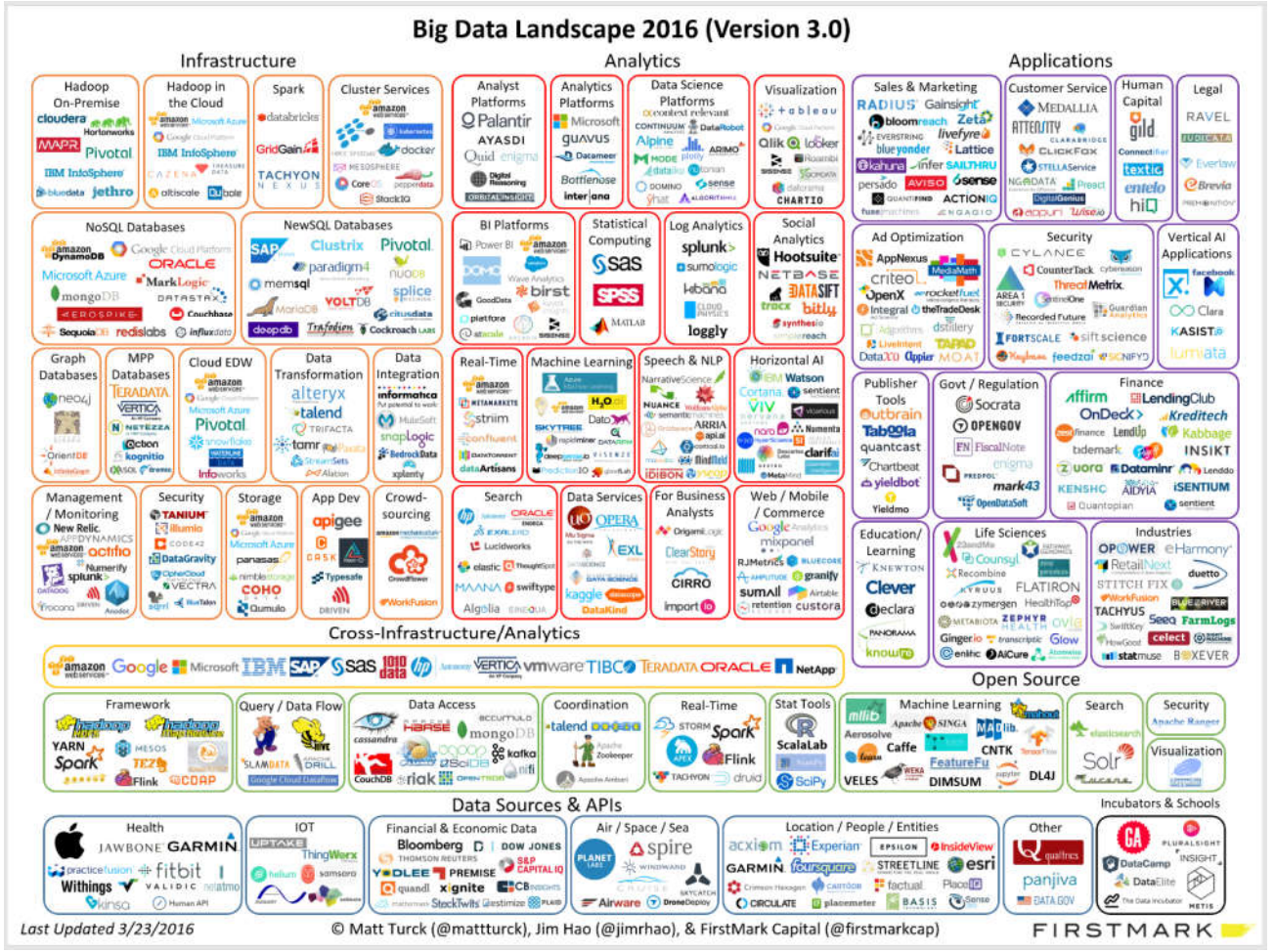
Sumber: https://www.datasciencecentral.com/profiles/blogs/data-scientists-data-engineers-software-engineers-the-difference

Differences



Sumber: https://www.springboard.com/blog/data-engineer-vs-data-scientist

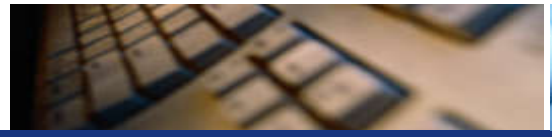
Data Engineering Tools



Sumber: https://joviam.com/this-infographic-of-big-data-tools-will-blow-your-mind-infographic/

Data Engineering Tools

Sumber: https://datafloq.com/big-data-open-source-tools/os-home/

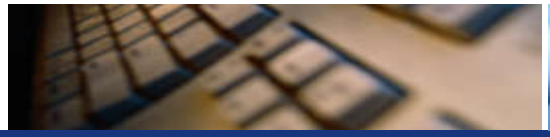


Database and Data Warehousing

Data Warehousing is used for reporting and data analysis. Big data however, requires different data warehouses than the traditional standard ones used in the past 10-20 years. There are multiple open source data warehouses available for different purposes:

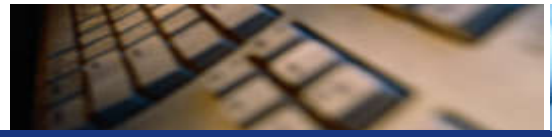
- **Infobright**

  Infobright offers a data warehouse that is scalable and that can store up to 50 terabyte of data. They offer a data compression technique that is up to 40:1 for better functioning. Next to open source do they also offer commercial products based on the same technology. It is especially designed to analyse large amounts of machine-generated data. The latest edition has the capabilities of nearly real-time analysis.




Database and Data Warehousing

- **Cassandra**
  Cassandra is a NoSQL database that was initially created by Facebook. The Apache Foundation however manages it today. The database is mainly used by large organisations that have massive active databases. Companies such as Twitter, Cisco, Netflix use it to optimize their databases. Cassandra also offers commercial support and services.
- **Apache Hbase**
  HBase is another Apache product and it includes linear and modular scalability. It is the non-relational data store for Hadoop. HBases is used by companies who need to random, real-time read/write access to Big Data. The objective of HBase is to host very large tables (billions of rows x millions of columns), while using commodity hardware.



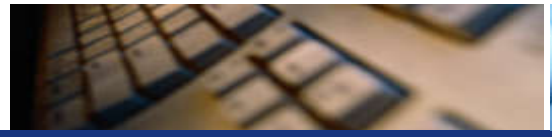

Database and Data Warehousing

- **Riak**

  Riak is a distributed database that is open source, scalable, fault-tolerant. IT is especially architected for replicating and retrieving data intelligently in order to read and write operations, even when the operations fail. Users can even lose access to nodes without losing data. Riak's customers are among others the Danish government, Boeing and Kipp.me.

- **Infinispan**

  Infinispan is a Java based data grid platform that was designed for multi-core architecture. It provides distributed cache capabilities. The objective is to have parallel data structures make the most of multi-core and multi processor architecture. Infinspan is not only available for Java, but also for PHP, Python, Ruby, C, etc. Infinispan also offer a ReST API to make connections with other websites easy going




Database and Data Warehousing

- **Bigdata**

  Bigdata is a distributed database that can scale from a single system to 1000s of machines. It is a horizontally scaled storage and the architecture provides for data-intensive, high performance distributed computing on commodity clusters. It can cope with many different data models, applications or workloads.

- **Hypertable**

  Hypertable is a NoSQL databse that allows fast performance and is efficient in use. It runs on top of a distributed file system such as Hadoop DFS and is written almost completely in C++. It offers comprehensive language support for languages such as Java, PHP, Python, Perl, Ruby etc. Although the software is completely open source, they also offer paid support.


Database and Data Warehousing

- **Terrastore**

  Terrastore is a document store providing advanced scalability. It relies on industry proven clustering technology (Terracotta) and it can be accessed via the HTTP protocol. It supports event processing, range queries, data partitioning, Mapreduce quering and processing functions. All queries and updates are distributed to the nodes thereby minimizing traffic and spreading computational load.

- **Apache Hive**

  This is Hadoop's data warehouse and it uses a SQL-like language called HiveQL. It promises easy data summarization, ad-hoc queries and other analysis of big data. It uses a mechanism to project structure onto data, while allowing map/reduce programmers to plug in custom mapper and reduces. Hive is an open source volunteer project under the Apache Software Foundation.


Database and Data Warehousing

- **Globals**
  Globals is the free database developed by InterSystems. It is a fast and scalable database offering multi-dimensional array storage. It is a NoSQL offering and it also offers an API that gives a rich approach to data modelling. The API is easy to use and fully adaptable. Globals is used by hundreds of thousands of sites.
- **Firebird**
  Firebird is a relational database that can run on Linux, Windows & various UNIX platforms. It offers high performance and powerful language support for stored procedures and triggers.
- **Oracle Berkely DB**
  The Oracle Berkeley DB family of open source, embeddable databases provides developers with fast, reliable, local persistence with zero administration. Berkeley DB enables the development of custom data management solutions, without the overhead traditionally associated with such custom projects.


Database and Data Warehousing

- **MariaDB**
  MariaDB is a backward compatible, drop-in replacement branch of the MySQL® Database Server. It includes all major open source storage engines + the Maria storage engine.
- **H2**
  H2 is a very fast Java SQL database with embedded and server modes, in-memory databases and a browser based console application. It has a very small footprint of only 1.5 MB
- **HyperSQL**
  It is a SQL relational database engine written in Java. HyperSQL offers a small & fast database engine which has in-memory and disk-based tables, supports embedded/server modes. Also, it has tools such as a command line SQL tool & GUI query apps.


Database and Data Warehousing

- **Drizzle**
  The Drizzle project is building a database optimized for Cloud and Net applications. It is being designed for massive concurrency on modern multi-cpu/core architecture. The code is originally derived from MySQL.
- **MonetDB**
  MonetDB is an open source column-oriented database management system developed at the Centrum Wiskunde & Informatica (CWI) in the Netherlands. It is a database system for high-performance applications in data mining, OLAP, GIS, XML Query, text & multimedia retrieval.
- **SQLite**
  SQLite is a software library that implements a self-contained, serverless, zero-configuration, transactional SQL database engine. SQLite is the most widely deployed SQL database engine in the world.
- **RethinkDB**
  RethinkDB is built to store JSON documents, and scale to multiple machines with very little effort. It has a pleasant query language that supports really useful queries like table joins and group by, and is easy to setup and learn.


Data Analytic Platform

There are several tools available which effectively are a Data as a Platform tool. These tools allow data analytics to be performed as a complete package.

- **Hadoop**
  Hadoop is the most well-known big data open source tool around at the moment. It supports data-intensive distributed applications that can run simultaneously on large clusters of normal, commodity, hardware. It is licensed under the Apache v2 license. A Hadoop network is reliable and extremely scalable and it works according to the computational model MapReduce. Hadoop is written in the Java programming language and is used by a global community of distributors.


Data Analytic Platform

- **Apache Spark**
  Apache Spark is an open-source cluster computing framework originally developed in the AMPLab at UC Berkeley. It is easy to use for developers, who can write applications in Java, Python or Scala. Programs run up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk. Spark comes with several libraries: Spark SQL, Spark Streaming, the MLlib machine learning library, and GraphX. It is scalable to 1000s of nodes and fault-tolerant.
- **Storm**
  Storm, which is now owned by Twitter, is a real-time distributed computation system. It works the same way as Hadoop provides batch processing as it uses a set of general primitives for performing real-time analyses. Storm is easy to use and it works with any programming language. It is very scalable and fault-tolerant.


Data Analytic Platform

- **MapReduce**
  MapReduce was originally developed by Google but has now been adapted by many big data tools, among others Hadoop. It is a software framework and model that can process vast amounts of data parallel on a large system of different computer nodes. The MapReduce libraries have been written in many programming languages and it therefore can work with all of them. MapReduce can work with structured and unstructured data.
- **HPCC Systems**
  HPCC means 'high performance computing cluster' and was developed by LexisNexis Risk Solutions. It is a similar version of Hadoop, but it claims to offer 'superior performance'. There is a free and paid version available. It works with structured and unstructured data and it is scalable from 1-1000s of nodes. It therefore also offers high-performance, parallel big data processing.


Data Analytic Platform

- **Hortonworks**
  Hortonworks is a pure open source Hadoop Distribution system. It is built on top of Hadoop and it allows users to capture, process and share data at any scale and in any format in a simple and cost-effective manner. Apache Hadoop is a core component of the Hortonworks architecture.
- **Dremel**
  Dremel is an interactive ad-hoc query system, which is developed by Google. IT offers analyses of read-only nested data. The system is extremely scalable; to 1000s of PCs and petabytes of data. It can process a collection of queries over massive, trillion-row, tables in just a matter of seconds by combining multi-level execution trees and columnar data layout.

Data Analytic Platform

- **Apache Drill**
  Apache Drill is part of the Apache Incubator and it offers a distributed system to perform interactive analyses of large-scale datasets that are based on Dremel. At the moment it is still incubating but the goals is to eventually become a massive scalable platform that can process petabytes of data in seconds over up to 10.000 servers.

- **GreenplumHD**
  Greenplum HD allows users to start with big data analytics without the need to built an entire new project. Greenplum HD is offered as software or can be used in a pre-configured Data Computing Appliance Module. IT exists of a complete data analysis platform and it combines Hadoop and Greenplum database into a single Data Computing Appliance.

Data Analytic Platform

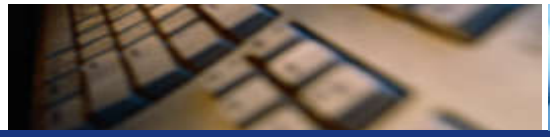
- **SAMOA**
  SAMOA is a platform for mining on big data streams. It is a distributed streaming machine learning (ML) framework that contains a programing abstraction for distributed streaming ML algorithms.

- **IKANOW**
  Ikanow focuses on developing products to enable uninhibited fusion and analysis of Big Data using open source technology. They have created an open source analytics platform.


In-Memory Open Source Tools

With the increasing amounts of data that need to be processed in real time, in-memory is gaining traction. Here are a few In-memory open source tools:

- **Gora**
  The Apache Gora open source framework provides an in-memory data model and persistence for big data. Gora supports persisting to column stores, key value stores, document stores and RDBMSs, and analyzing the data with extensive Apache Hadoop™ MapReduce support.
- **GridGrain**
  GridGain is an enterprise open source grid computing made for Java. It is compatible with Hadoop DFS and it offers a substitute to Hadoop's MapReduce. GridGain offers a distributed, in-memory and scalable data grid, which is the link between data sources and different applications. An open source version is available on Github or a commercial version can be downloaded from their homepage.




In-Memory Open Source Tools

- **Hazelcast**
  Hazelcast is the open source clustering and very scalable data distribution platform for Java. It allows users to partition and share data across a distributed network. It is a peer-to-peer solution and therefore does not require a master node. It is simple to use and comes with an web-based cluster monitoring tool.

- **Nmemory**
  NMemory is a lightweight non-persistent in-memory relational database engine that is purely written in C#. It can be hosted by .NET applications and supports traditional database features like indexes, foreign key relations, transaction handling and isolation, stored procedures, query optimization, field constraints.


In-Memory Open Source Tools

- **Terracotta**

  Terracotta is an in-memory data management platform that is built on top of Ehcache. Ehcache is the de facto Java caching solution that has been a favorite of the open source community since 2003. Terracotta also offers Quartz, a full-featured, open source job scheduling service that can be used with any Java application. Terracotta server is another offering for the open source community that supports fail over for in-memory data. These three offerings make up the Terracotta open source community that is 2+ million member strong and growing.

Business Intelligence Tools

Business Intelligence gives organisations the possibility to collect, maintain and organize data in order to identify new (business) opportunities. BI can help while implementing an effective strategy. More and more open source tools appear that offer business intelligence beyond the standard platforms like Google Analytics.

- **Talend**
  The open source software developed by Talend has developed several big data software solutions, including Talend Open Studio for Big Data, which is a data integration tool supporting Hadoop, HDFS, Hive, Hbase and Pig. The objective is to improve the efficiency of data integration job design through a graphical development environment. Next to open source tools, Talend also sells other commercial products.


Business Intelligence Tools

- **Jaspersoft**

  Jaspersoft has developed several open source tools, among others a Reporting and Analytics server, which is a standalone and embeddable reporting server. The Open Source Java Reporting Library is a reporting engine that can analyse any kind of data and produce reports in any format. Jaspersoft ETL offers a data integration engine, powered by Talend. They claim it is the world's most used business intelligence software.

- **SpagoBI**

  SpagoBI is an open source business intelligence suite. It provides analytical capabilities that are supported by over 30 analytical and operational engines. It is very flexible and an easy deploying into existing IT infrastructure. The suite also includes solutions for innovative intelligence domain such as KPI's, visual enquiries, real time BI and mobile BI.


Business Intelligence Tools

- **Pentaho**
  Open source tool Pentaho is a big data tool focusing on data mining and reporting in order to create complex solutions for complex problems. Pentaho offers also paid products, but the open source tool is supported by a large community dedicated to delivering a complete, well integrated, and high quality suite of business intelligence software.
- **BIRT Exchange**
  BIRT stands for "Business Intelligence and Reporting Tools" and it is an Eclipse-based tool, which adds reporting features to Java applications. IT can produce compelling reports. The objective is to combine a wide-range of reporting needs within one application. It has two components: 1) a visual report designer and 2) a runtime component for creating reports that can be deployed to any Java environment.
- **OpenI**
  OpenI offers open source dashboards, interactive reporting, ETL, predictive modeling and more. Operating System: OS Independent. OpenI is a web-based OLAP reporting application. OpenI is an out-of-box solution for building and publishing reports from XMLA-compliant OLAP data sources.


Data Mining Tools

Data mining is the process of discovering patterns in massive amounts of data. Methods like machine-learning, database systems, artificial intelligence and statistics can help to extract information from the data in an understandable way in order to use it for business decisions.

- **RapidMiner**
  RapidMiner offers data integration and analysis, analytical ETL and reporting combined in a community edition or enterprise edition. It comes with a graphical user interface for designing analysis processes. The solution offers a meta data transformation, which allows inspecting for errors during design time.


Data Mining Tools

- **KNIME**
  KNIME [naim] is a user-friendly graphical workbench for the entire analysis process. The Konstanz Information Miner offers data access, initial investigation, data transformation, predictive analyses and reporting tool with a user-friendly graphical interface. They offer an open-source tool and commercial products. The open integration platform offers over 1.000 modules. KNIME is also the open source data analytics platform continuously ranked No. 1 in customer satisfaction.
- **Mahout**
  Mahout is another project by Apache. It offers algorithms to build machine-learning libraries that can scale to reasonable large data sets. It is especially made for classification, clustering and batch-base collaborative filtering that run on Hadoop. Non-Hadoop and single-node contribution can also be used and the core libraries are optimized also for non-distributed algorithms.


Data Mining Tools

- **Orange**
  As the name indicates, it is an open source tool aiming to make data mining "fruitful and fun". It offers data mining through the use of visuals. It is data visualizations and analysis made easy for novice users as well as for experts. User can design data analyses through visual programming and Python scripting.

- **WEKA**
  WEKA stands for "Waikato Environment for Knowledge Analysis" and it is a collection of machine-learning algorithms in order to solve data mining problems. It is written in Java and thus runs on almost any modern computing platform. It supports different data mining tasks such as clustering, data pre-processing, regression, classification, feature selection as well as visualization.

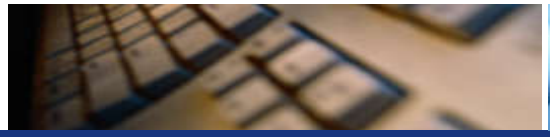

Data Mining Tools

- **KEEL**
  KEEL is an abbreviation for Knowledge Extraction based on Evolutionary Learning. It is based on Java and it can assess the behaviour of evolutionary algorithms used for data mining related problems, among others clustering, regression or classification. KEEL was also designed for research and educational purposes.
- **Togaware**
  Togaware has developed the "R Analytical Tool To Learn Easy" also known as Rattle. This is a graphical user interface for data mining using the R language. It can present visual and statistical summaries of data. It can built forms out of data that are easily modelled, present the performance of models graphically and it can score new datasets.
- **SPMF**
  Also a Java based data mining. It includes 51 different algorithms among others for sequential pattern and or rule mining, frequent itemset mining, clustering and association rule mining. The source code of each algorithm can be combined with other Java programs. It can be used with an interface or from the command line.


Object Database

Object databases first started to appear in 1985 and they are also known as object-oriented database management systems. The data in such databases is stored in the form of objects, as used by object-oriented programming. They are different from relational or graph databases and most of them offer a query language that allows object to be found with a declarative programming approach.

- **Db4objects**
  Db4o is the open source object database that is supported by a large community. The database allows .Net and Java and developers to store and recover any application object with just 1 line of code. Versant, the developer of Db4o offers a free and commercialized suite.
- **PicoLisp**
  PicoLisp is a Lisp dialect running on Linux and it features a database functionality. First class objects are loaded from database files automatically when they are accessed and written back when modified. PicoLisp was developed in 1988 and in 2012 a Java version was created.




Object Database

- **FramerD**
  FramerD, developed in 2005, allows computerized creation, access and manipulation of descriptions and systems of descriptions. It is a portable distributed object-oriented database and was developed to support sharing and maintaining knowledge bases. It is optimized for pointer-intensive data structure.

- **Siaqodb**
  Siaqodb is a NoSQL object database running on among others .NET, Unity3D as well as the Windows Phone. Thanks to the Sync Framework provider, siaqodb is a cross-platform client-side database. It can be kept in sync with server-side databases. LINQ is the query engine and it provides a LINQ query editor in the application.

Object Database

- **McObject**
  McObject developed Perst, an open source object-oriented embedded database sytem (ODBMS). Data is stored in Java directly and it is available for .NET framework applications as well. Users can store, sort and retrieve objects in high-speed and with memory usages being low. It is reliable to use and offers different development tools.

- **Starcounter**
  Starcounter is a transactional database made available for modern commodity computers. Transactions are secured on disk and it supports duplication and full recovery. It integrates Virtual Machine and a Database Management System. It offers a .NET object API and SQL query support and gives a single server the capacity of a data centre.

Object Database

- **Zope**
  Zope is developed and maintained by one of the largest open source communities. It was developed in 1998 in the object-oriented programming language Python. Zope stands for 'Z Object Publishing Environment' and Zope is helped the programming language Python become popular. It was the first system to start using object publishing methodology for the web.

- **Magma**
  Magma is a multi-user object database developed especially for Squeak 4.4. It provides good read-transparency to a large-scale shared persistent object model. It has a high-availability and fault tolerance and is developed to support large indexed collections with hard querying.


Object Database

- **Ndatabase**
  NDatabase is a .NET Object Database and a transparent persistence layer for .NET. It allows users to store and retrieve native objects with a single line of code. It offers NoSQL and LINQ support and databases are automatic created. It supports different platforms among other Silverlight, Windows Phone and NuGet.
- **Neoppod**
  Neoppod is a distributed, redundant and transactional storage system. It was initiated in 2005 and implemented on top of Zope. It is a NoSQL database developed for the cloud.
- **Sterling**
  Sterling is a NoSQL object-oriented database developed especially for Silverlight, Windows Phone 7.0 and .NET. It supports LINQ object queries. The core is light so that the system is flexible and it becomes easy to query the database. Sterling is portable and weighs only 85 kb.


Object Database

- **EyeDB**
  EyeDB is an Object Oriented Database Management System (OODBMS) that is developed by SYSRA. It provides an object model, object definition language, a manipulation language and an object query. It offers programming interfaces for C++ and Java.

- **Persevere**
  Persevere is an object storage engine and application server (running on Java/Rhino) that provides storage of dynamic JSON data for rapidly develop data-driven JavaScript-based rich internet applications


XML Database

XML Databases allow data to be stored in XML format. XML databases are often linked to document-oriented databases. The data stored in an XML database can be queried, exported and serialized into any format needed. There are two different types of XML databases:

- XML-enabled: these databases can map XML to e.g. a relational database, accept XML input or render XML as output.
- Native XML: it uses XML documents as the fundamental unit of storage.

- **Existdb**
  Exist-db is an open source native XML database. It supports many web 2.0 technology standards and it is therefore a suitable platform for web-based applications. It has an HTTP interface and features index-based XQuery processing. It requires Java to operate.
- **BaseX**
  BaseX is a scalable XML database that is light-weight and offers high-performance. It uses a Graphical User Interface for the front end to give users insight into the stored XML documents. It supports very large XML documents and a real-time XQuery editor.

XML Database

- **Qizx**
  Qizx is an XQuery processor that has been open source since 2003. Is does not have a persistent storage and XML documents have to be analysed in memory before they can be used. It offers average sized XML documents a fast query. Another aspect of Qizx is that it is also an XML database that offers storage and indexing capabilities.

- **Sedna**
  Sedna is a native XML database that offers a complete range of principal datastore services. It offers full-text search and persistent storage as well as flexible XML processing facilities. It is licensed under the Apache License and has external XQuery functionalities implemented in C.

XML Database

- **Xindice**
  Apache Xindice is a retired database that was designed from the ground up to store XML data. With Xindice data can be inserted and retrieved as XML. It works well with very complex XML structures that normally require a more structured database.

- **LiquieBase**
  Liquibase is an open source database-independent library for tracking, managing and applying database changes. It has over 30 built-in database refactorings and users can make custom changes.


Graph Database

Graph databases use graph structures (a finite set of ordered pairs or certain entities), with edges, properties and nodes for data storage. It provides index-free adjacency, meaning that every element is directly linked to its neighbour element. No index lookups are necessary. Graph database are faster when it comes to associative data set compared to relational databases. As they do not need join operations, they can scale naturally to large data sets.

- **Gephi**
  Gephi helps people understand and discover graphs and patterns. It uses a 3D engine to show graphs in real-time that can help users make hypothesis, isolate structure singularities or fault during data sourcing. It is written in Java on the Netbeans platform. It can be used to analyse graphs extracted from OrientDB.

- **FlockDB**
  FlockDB is a simple graph database and is intended for online low-latency, high throughput environments, like websites. FlockDB is being used by Twitter to store social graphs. It is a distributed graph database and can support complex arithmetic queries. The database is licensed under the Apache License.


Graph Database

- **GraphBuilder**
  GraphBuilder can reveal hidden structures in big data as it can construct graphs out of large sets of data. It is developed by IBM, built in Java, it uses Hadoop and it scales using the Map Reduce parallel processing model. The GraphBuilder library takes care of many of the difficulties of graph construction, such as graph transformation, formation and compression.
- **InfoGrid**
  InfoGrid is developed in Java and at the heart of it lies the GraphDatabase. It offers many additional software components that makes it easy to develop graph based web applications. InfoGrid is sponsored by NetMesh and they offer also commercial support for using InfoGrid.
- **InfiniteGraph**
  InfiniteGraph helps users to ask more complex and deeper questions across their data stores. It can work with massive amounts of distributed data and especially those projects that need more than one server will benefit the most from the graph database. It offers high-speed graph traversals, scalability and parallel consumption of the data.


Graph Database

- **Franz Inc – AllegroGraph 4.9**
  AllegroGraph is a graph database with MongoDB integration. It is developed for high-speed and high-performance loading and query speed. It uses effective memory utilization and it can scale to massive amounts of quads. It supports SPARQL and RDFS++ and it has a Javascript-based interface.
- **Gremlin**
  Gremlin is a graph traversal language and it can be used for graph analysis, query and manipulation. Gremlin works with the graph databases that have included the Blueprints property graph data model. These include among others Neo4j, OrientDB, InfiniteGraph. Gremlin provides native support for Java and Groovy.
- **HypergraphDB**
  HyperGraphDB is a general purpose data storage mechanism designed for knowledge representation. It is based on directed hypergraphs and offers graph-oriented storage. It can be used as an embedded object-oriented database for Java projects or as a NoSQL relational database. The core of the database engine is designed for generalized, typed and directed hypergraphs.


Graph Database

- **GraphBase**
  GraphBase is a Graph Database Management System that was built from scratch in order to manage large graphs. It makes huge, very structured data stores possible. Graphbase simplifies the usage of graph-structured data, instead of working with very complex spaghetti-like structures. With GraphBase Singleview, it becomes possible to turn a database into a single, searchable and navigable graph.

- **BrightstarDB**
  Brightstar DB Mobile and Embedded are the open-source tools of BrightstarDB. It is a NoSQL database designed for the .NET platform that is fast, embeddable and scalable. It does not need fixed schema, which gives is a lot of flexibility in what and how the data is stored. Its associative data model fits perfectly with real world applications.


Graph Database

- **Sparksee**
  Sparksee (formerly known as DEX), makes space and performance compatible with a small footprint and a fast analysis of large networks. It is natively available for .Net, C++, Python and Java, and covers the whole spectrum of Operating Systems. Sparksee mobile is the first graph database available for iOS and Android.

- **Neo4j**
  Neo4j is a graph database boasting massive performance improvements versus relational databases. It is very agile and fast. At the moment it is used by many startups in applications such as social platforms, fraud detection, recommendation engines etc. The data is stored in nodes that are connected by directed, typed relationships with properties of both (a property graph).

SELESAI