Appendix A
Nomenclature for Managers

Relational database management system (RDBMS): structured data in predetermined schema (*tables*), scalable vertically through large SMP servers, or horizontally through clustering software. These databases are usually easy to create, access, and extend. The standard language for relational database interoperability is the **Structured Query Language (SQL)**.

Non-relational database: database that does not store data into tables, but made them accessible through special query APIs. The standard language used is **Not Only SQL (NoSQL)**: it does not present a fixed schema, it uses BASE system to scale vertically (basically available, soft-state, eventually consistent), and sharding (horizontal partitioning) to scale horizontally. Examples are **MongoDB** and **CouchDB** (they differ mainly because in MongoDB the main objects are documents, while in CouchDB are collections, which in turn contain documents). NoSQL commonly used **JavaScript Object Notation (JSON)** data format (**BSON** in MongoDB—binary JSON), and it mainly works through **Key Value Store (KSV)**, i.e., a collection of different unknown data types (while a RDBMS stores data into table knowing exactly the data type).

Hadoop: open source software for analyzing huge amount of data on a distributed system. His primary storage is called Hadoop distributed file system (**HDFS**), which duplicates the data and allocates them in different nodes. It has been written in Java. It is a core technology in the big data revolution and stores data into their native raw format, and it can be used for several purposes (Dull, 2014), such as a simply data staging or landing platform complementary to the existing EDW (as an enterprise data hub, i.e., EDH), or managing data (even small), transforming those into a specific format in the HDFS and sending them back to the EDW, lowering thus the costs while increasing the processing power. Furthermore, it can integrate external data-sources and archive data (both on-premises or into the cloud), and reduce the burden for a standard EDW.

MapReduce: software for parallel processing huge amount of data.

© Springer Nature Switzerland AG 2019

F. Corea, *An Introduction to Data*, Studies in Big Data 50,
https://doi.org/10.1007/978-3-030-04468-8

Flume: service to gather, aggregate, and move chunks of data from several sources to a centralized system.

Cassandra: an open source database system for analyzing large amount of data on a distributed system. It is characterized by a high performance and by a high availability with no single point of failure (i.e., a part of system that if fails stop the whole system). It fosters data denormalization, which means grouping data or adding redundant information, in order to optimize the database performance.

Distributed System: Multiple terminals communicating between them. The problem is divided in many tasks, and assigned to each terminal. It is a highly scalable system as further nodes are added.

Google File System: proprietary distributed file system for managing efficiently large datasets.

HBase: an open source non-relational database (column-oriented) developed on a HDFS. It is very useful for real time random read and write access to data, as well as to store sparse data (small specific chunk of data within a vast amount of them). The relational counterpart is called **Big Table**.

Enterprise Data Warehouse (EDW): system used for analysis and reporting that consists of central repositories of integrated data from a wide spectrum of different sources. The typical form of an EDW is the **extract-transform-load (ETL)**, that is the most representative case of *bulk data movement*, but other three important examples of these systems are **data marts** (i.e., a subset of the EDW extracted out in order to address a specific question), **Online analytical processing (OLAP)**—used for multidimensional low-frequency analytical query—and **Online transaction processing (OLTP)**—used rather for high volume fast transactional data processing. The wider system that includes instead a set of servers, storage, operating systems, database, business intelligence, data mining, etc. is called **data warehouse appliance (DWA)**.

Resilient Distributed Datasets (RDD): logical collection of data partitioned across machines. The most known examples is **Spark**, an open source clustering computing that has been designed to accelerate analytics on Hadoop thanks to the multi-stage in-memory primitives (that are basic data types defined in programming languages or built it with their support). It seems to run 100 times faster than Hadoop, but its disadvantage is that it does not provide its own distributed storage system.

Hive: additional example of EDW infrastructure that facilitates data summarization, ad-hoc queries, and specific analysis.

Pig: platform for processing huge amount of data through a native programming language called Pig Latin. It runs at the same time sequences of MapReduce.

Programming language: is a formal constructed language designed to communicate instructions to a machine. The main ones for data science applications are Java, C, C++, C#, R, and Matlab. Scala is another language that is becoming extremely popular right now.

Scripting Language: is a programming language that supports scripts, which are piece of codes written for a run-time environment that interpret (rather than

compile) and automate the execution of tasks. The main ones in big data field are Python, JavaScript, PHP, Perl, Rub, and Visual Basic Script.

Data Mart: is a subset of the data warehouse used for a specific purpose. Data marts are then department-specific or related to a single line of business (LoB). The next level of data marts is the **Virtual Data Marts**, i.e., a virtual layer that create various views of data slices—in other words, instead of physically creating a data mart, it just takes a snapshot of them. The final evolution is instead called **Data Lakes**, which are massive repositories of unstructured data with an incredible computational capability. Hence, data marts physically create repositories (slices) of data, virtual data marts leave the data where they are and create virtual constructs—reducing the cost of transferring and replicating them—while data lakes work as the virtual data marts but with any kind of data format.

Appendix B
Data Science Maturity Test

The following questionnaire provided could help managers to grasp a rough idea of the current data stage of maturity they are facing within their organizations. It has to be integrated with deep conversations and meetings with the big data analytics (BDA) staff, the IT team, and supported by solid researches.

(1) What is your investment level in BDA capabilities?

1. Absent. We don't have money for big data
2. A small budget is allocated when positive quarters in core activities allow us to do that
3. A modest funding scheme is in place
4. We invested a good percentage of our revenues in BDA in the last year, and we will keep investing because it is part of our company's vision.

(2) What the executives' support to analytics capabilities?

1. Neither IT nor business think BDA is useful to the business
2. Only IT managers support it because they are interested in the technological challenge
3. Business managers see the hidden value in data and support BDA projects
4. Both IT and business executives believe in BDA potential.

(3) What is your current stage of working with data?

1. We will start using data in the future if needed
2. We have a good idea of what business questions we could solve with data in my company
3. We take action using analytics

© Springer Nature Switzerland AG 2019

F. Corea, *An Introduction to Data*, Studies in Big Data 50,
https://doi.org/10.1007/978-3-030-04468-8

4. We are automating analytics the most we can, and we believe is a competitive factor that gives us benefits we are able to communicate frequently to top management and shareholders.

(4) Your analytics team is:

1. Inexistent
2. Acquired from outside at the moment
3. We have some senior scientist that has been recruited, but we are now growing the team internally by training
4. An independent sustainable group and function within the company.

(5) Your company's culture is:

1. Intolerant—especially for failure concerning new analytics, methodologies, and technologies
2. Variegated—it is half-half made by old-style professionals and geeks
3. Collaborative—people are willing to work together and share.
4. Creative—innovation is valued and we are encouraged and monetarily compensated for our original shared contributions.

(6) How your data science team is connected to the company hierarchy?

1. We only have some analysts with small tasks, who deliver the outcomes to their direct managers on a weekly/monthly basis
2. The data team is leaded by a business head, and their contribution is continuously marginally positive
3. Our data scientists are tight to our data warehouse and data management teams, and they constantly interconnected with the business side
4. They are autonomous and do not seat in the same building of the operations function. They are allocated in a Centre of Excellence.

(7) The internal data policy is:

1. Fairly poor, we do not need it
2. Metadata definitions and BDA policy are well-established
3. We have a BDA policy that we constantly monitor and we have a security policy for any data forms
4. We have a BDA and security policies, and we anonymized all the relevant data to protect our clients and partners' privacy.

(8) The data in your company are:

1. Stored in silos
2. We prioritized the data to be used within our organization, and they are internally shared

3. Many different data sources are integrated for our analysis, and we take care of data quality through a meticulous goodness assessment based either on the final use or the type of data we will exploit
4. We have integrated BDA technologies into our systems, we store our data on a cloud, and we often use them for mobile applications.

(9) When your company looks at your BDA capabilities:

1. It sees mainly a sunk-cost, i.e., the cost of storing, maintaining, protecting and analyzing these datasets
2. We know data have value and we understand both the data cost and data competitive advantage, but we are definitely overwhelmed
3. We are rationalizing our data storage and usage abilities, because we understood that not everything is either pertinent or meaningful
4. We have an efficient process for data aggregation, integration, normalization and analysis, and we can manage easily any amount of inflowing data.

(10) Your firm is currently using:

1. Relational Database and Internal data
2. Data marts, R or Python languages, and public data
3. NoSQL database, Hadoop and MapReduce, and we use external data, sometimes also unstructured
4. Highly unstructured data, APIs, and a Resilient Database.

Once each single question has been answered, it is simple to obtain a rough measure of the data maturity stage for a certain company. For each answer indeed, it has to be considered the number associated to that answer, and then it is enough to sum up all the numbers obtained in this way. So, for example, if in the third question the answer is "we take action using analytics", the number to be considered is 3, since it is the third answer of the list.

Finally, the score obtained should range between 1 and 40. The company will then belong to one of the four stages explained in Table 2.1 accordingly to the score achieved, that is explained in the Table B.1:

Table B.1 Data science maturity test classification

	Primitive	Bespoke	Factory	Scientific
Score	10–15	16–25	26–35	36–40

Appendix C
Data Scientist Extended Skills List (Examples in Parentheses)

Programming (R, Python, Scala, JavaScript, Java, Ruby, C++).

Statistics and Econometrics (probability theory, ANOVA, MLE, regressions, time series, spatial statistics).

Bayesian Statistics (MCMC, Gibbs sampling, MH Algorithm, Hidden Markov Model).

Machine Learning (supervised and unsupervised learning, CART).

Mathematics (Matrix algebra, relational algebra, calculus).

Big Data Platforms (Hadoop, Map/Reduce, Hive, Pig, Spark).

Text mining (Natural Language Processing, SVM, LDA, LSA).

Visualization (graph analysis, social/Bayes/neural networks, Tableau, ggplot, D3, Gephi, Neo4j).

Business (business and product development, budgeting and funding, project management, marketing surveys).

Algorithms (SVM, PCA, GMM, K-means, Deep Learning).

Optimization (linear, integer, convex, global).

Simulations (Monte Carlo, agent-based modeling, NetLogo).

Structured Dataset (SQL, JSON, BigTable).

Unstructured Dataset (text, audio, video, BSON, noSQL, MongoDB, CouchDB).

Multi-structured Dataset (IoT, M2M).

Data Analysis (feature extraction, stratified sampling, data integration, normalization, web scraping).

Systems Architecture and Administration (DBA, SAN, cloud, Apache, RDBMS).

Scientific approach (experimental design, A/B testing, technical writing skills, RCT).

© Springer Nature Switzerland AG 2019

F. Corea, *An Introduction to Data*, Studies in Big Data 50,
https://doi.org/10.1007/978-3-030-04468-8

Appendix D
Data Scientist Personality Questionnaire

The terminology used to classify into 16 subcategories the different kind of data scientists is given by the two-entry matrix exhibited in the Table 7.1. The terminology can be sometime misleading if related to the Keirsey Temperament Sorter (KTS), and this is why it is necessary to specify that the only categorization borrowed from KTS framework is the broader one, i.e. the Artisan-Idealist-Rational-Guardian partition. Every sub-category has instead to be taken as newly generated.

Here it follows the personality test to sort data scientists into a specific box. It is composed by 10 questions, and for each one a single answer has to be provided. This test is not a professional temperament test to fully understand individuals' personality, but it is more a quick tool for managers to efficiently and consciously allocate the right people to the right team.

(1) When you start working on a new dataset

   a. You start exploring immediately and querying the data
   b. Plan in advance how to tackle it
   c. You spent time in understanding the data, where they come from, and their meaning
   d. You identify a research question quickly, and focus on designing the a new improved method for analyzing your data.

(2) In your team, people count on you for your

   a. Troubleshooting ability
   b. Organizational skills
   c. Capacity to reduce the problem complexity
   d. Strategic approach and conceptualization of the problem.

© Springer Nature Switzerland AG 2019

F. Corea, *An Introduction to Data*, Studies in Big Data 50,
https://doi.org/10.1007/978-3-030-04468-8

(3) When facing a new data challenge, your first thought is

a. Is what I am doing impactful and relevant?
b. When do I have to deliver some results?
c. How this challenge can make me better?
d. What I can learn from this dataset?

(4) In a data analysis, which is the most important thing to you

a. Results, no matter how you do achieve them, what strategy or technology you do employ
b. To achieve a result in the correct way and with the right process or technology
c. Attaining significant results in an ethical manner
d. Reaching the outcomes through an accurate, replicable, and efficient procedure.

(5) If you have finished your assigned today's work, you would

a. Focus again on your analysis and try to find alternative and innovative way to achieve your final goal
b. Start with something else, even if this may mean to stay longer at your desk
c. Help a colleague in difficulty with his analysis
d. Give suggestions and highlight weaknesses in your colleagues' works for the sake of the team and business development.

(6) If you would have some spare time during your daily work, you would prefer to

e. Optimize existing technology for the whole company
f. Improve your analysis
g. Try to derive new insights from your previous analysis
h. Understanding how to maximize the value of your analysis.

(7) It is your data-dream of

e. Speaking about data with only engineers and IT team
f. Teaching data related contents
g. Engaging with people who do not know anything about data science
h. Persuading and convincing the business team of the big data opportunity

(8) You prefer to work with

e. Huge amount of structured data
f. Any kind of data that challenge me
g. Behavioral or social media data, or any unusual data
h. No data in particular.

(9) If you would quit tomorrow your data science job, you would prefer to become

e. An IT manager or software engineer
f. A professor
g. A consultant
h. An entrepreneur.

(10) What characteristic of big data you value the most

e. Volume
f. Velocity
g. Variety
h. Value.

Once each question has been answered with a single reply, the result is given by pairing the reply chosen more often within the first five questions (a–d) with the answer that appears more often in the last five (e–f), as shown in the following table. So, if for instance in the first five questions *b* emerges as predominant answer, while in the last five *f* is the median, the person considered is a *Cruncher* (Table D.1).

Table D.1 Data scientist personality classification

Archetype/ personality	Artisan	Guardian	Idealist	Rational
Technical	*Gardener*: A–E	*Architect*: B–E	*Evangelist*: C–E	*Wrangler*: D–E
Researcher	*Alchemist*: A–F	*Cruncher*: B–F	*Champion*: C–F	*Groundbreaker*: D–F
Creative	*Trailblazer*: A–G	*Catalyst*: B–G	*Visionary*: C–G	*Warlock*: D–G
Strategist	*Babelian*: A–H	*Mastermind*: B–H	*Advocate*: C–H	*Fisherman*: D–H