A Comparison of Data Science Systems

Carlos Ordonez

University of Houston, USA

Abstract. Data Science has subsumed Big Data Analytics as an interdisciplinary endeavor, where the analyst uses diverse programming languages, libraries and tools to integrate, explore and build mathematical models on data, in a broad sense. Nowadays, there exist many systems and approaches, which enable analysis on practically any kind of data: big or small, unstructured or structured, static or streaming, and so on. In this survey paper, we present the state of the art comparing the strengths and weaknesses of the most popular languages used today: Python, R and SQL. We attempt to provide a thorough overview: we cover all processing aspects going from data pre-processing and integration to final model deployment. We consider ease of of programming, flexibility, speed, memory limitations, ACID properties and parallel processing. We provide a unifying view of data storage mechanisms, data processing algorithms, external algorithms, memory management and optimizations used and adapted across multiple systems.

1 Introduction

The analytic revolution started with data mining [2], which adapted machine learning and pattern detection algorithms to work on large data sets. In data mining scalability implied fast processing beyond main memory limits, with parallel processing as a secondary aspect. Around the same time data warehousing [2] was born to analyze databases with demanding queries in so-called On-Line Analytical Processing [1]. During the last decade Data Warehouses evolved and went through a disruptive transformation to become Big Data Lakes, where data exploded in the three Vs: Volume, Velocity and Variety. The three Vs brought new challenges to data mining systems. So data mining and queries on Big Data lakes morphed into a catch-all term: Big Data Analytics [3]. This evolution brought faster approaches, tools and algorithms to load data, querying beyond SQL considering semantics, mixing text with tables, stream processing and computing machine learning models (broadly called AI). A more recent revolution brought two more Vs: Veracity in the presence of contradictory information and even Value, given so many development and tool options and the investment to exploit big data. Nevertheless, having so much information in a central repository enabled more sophisticated exploratory analysis, beyond multivariate statistics and queries. Over time people realized that managing so much diverse data required not only database technology, but also a more principled approach laying its foundation on one hand in mathematics (probability, machine learning, numerical optimization, statistics) and on the other hand, more abstract, highly

analytic, programming (combining multiple languages, pre-processing data, integrating diverse data sources), giving birth to Data Science (DS). This new trend is not another fad: data science is now considered a competing discipline to computer science and even applied mathematics.

In this survey article, we consider popular languages used in Data Science today: Python, R, and SQL. We discuss storage, data manipulation functions, processing and main analytical tasks across diverse systems including their runtimes, the Hadoop stack and DBMSs. We identify their strengths and weaknesses in a modern data science environment. Due to space limitations examples with source code and diagrams are omitted.

2 Infrastructure

2.1 Programming Language

The importance of the specific programming language used for analytic development cannot be overstated: it is the interface for the analyst, it provides different levels of abstraction, it offers different data manipulation mechanisms and it works best with some specific analytic task. There are two main families: interpreted and compiled. DS languages are interpreted, but the systems evaluating them have compiled languages behind, mostly C++ (and C), but also Java (Hadoop stack). Given the need to provide high performance the DS language offers extensibility mechanisms that allow plugging in functions and operators programmed in the compiled language, but which require expertise on the internals of the runtime system. The average analyst does not mess with the compiled language, but systems researchers and developers do. Finally, SQL is the established language to process queries in a database and to compute popular machine learning models and data mining techniques. In a similar manner to DS languages, it is feasible to extend the language with advanced anlytic capabilities with UDFs programmed in C++/C and a host language (Java, Python).

2.2 Storage Mechanisms

In this section we attempt to identify the main storage mechanisms used today: arrays, data frames and tables. Notice each programming language offers these mechanism is different flavors as well as with different capabilities and constraints.

An array is a data structure provided by a programming language and it is arrangement of elements of the same data type. An array can represent a table as an array of records and a matrix as a 2-dimensional array of numbers. In most programming languages, like C++ or Java, the array is a data structure in main memory, although there exist systems (SciDB [5]) which allow manipulating arrays on disk. The data frame is another data structure in main memory, somewhat similar to an array. Data frame was introduced in the R language, allowing to assemble rows or columns of diverse data types, which has made its

way into Python in the Pandas library. It has similarities to a relational table, but it is not a relational table. When the data type of all entries is a number data frames resemble a matrix, but they are not a matrix either. However, it easy to transform a data frame with numbers into a matrix and vice-versa. A table is a list of records, which can be stored as arrays in main memory or as blocks (small arrays) on disk. By adding a primary key constraint a generic table becomes a relational table [1]. A relational table is a data structure on secondary storage (disk, solid state), allowing manipulation roow by row, or block by block. A record is generally understood as physical storage of a tuple, followimg a specific order and each value taking fixed space. Thus, by design, a relational table is very different from a data frame. Data frames and arrays are somewhat similar, but data frames track row and column names and they do not require contiguous storage in main memory.

The I/O unit is an important performance consideration to process a large data set, which varies depending on the system: line by line or an entire file for a data frame. one line at a time for text files, one record or block of records for tables. DS systems and libraries vary widely on how they read, load and process data sets. Arrays vary according to their content; in 2-dimensional format of numbers they are read in one pass, as one block for a small array or multiple blocks for a large array.

2.3 Physical Primitive Access Operators

We now attempt to identify the main access primitive operators for each storage mechanism.

All systems can process text and binary files, but tend to favor one. DS languages generally work on text files, which can be easily transferred and exchanged with other analytic tools (spreadsheets, editors, word processors). DBMSs use binary files in specific formats, doing a format conversion when records are inserted by transactions or when they are loaded in batch from files. Hadoop system use a combination of both, ranging from plain files to load data to transforming data into efficient block formats as needed in a subsystem.

Large arrays are generally manipulated with blocked access on binary files, where each block is an array. Blocks for dense matrices are generally squared. Few systems attempt to provide subscript-based access on disk because it requires adding special functions and constructs in the programming language.

Data frames are generally loaded with one scan on the input text file, but there exist libraries that allow block by block access (called chunk to distinguish them from database blocks). Therefore, the access is sequential.

Each programming language and system incorporate different access operators, but here we provide a broad classification. In a DS language these are the most common operators in main memory: scan (iterator), sort, merge. Iterators come from object-oriented programming. The merge operator is similar to a relational join, but more flexible to manipulate dirty data. DBMSs feature database operators combining processing in main memory and disk: scan, sort, join; filters are processed with a full scan when there are no indexes, but they

can be accelerated with indexes when possible. Sorting can be used at many stages to accelerate further processing.

3 Adding New Data and Updating Old Data

3.1 Input Data Format

It is either text or binary. The most common input file format are text files, with CSV format being a de facto standard. The CSV format allows representing matrices, relational tables and data frames with missing values and strings of varying length. Binary formats are more common in HPC and specific science systems (physics, chemistry, and so on). Streams come in the form of ever growing log files (commonly CSV), where each log record has a timestamp and records that can vary in structure and length. JSON, a new trend, is a more structured text format to import data, which can represent diverse objects, including documents and database records.

3.2 Copying Files vs Loading Data

In most data science languages there is no specific loading phase: any file can be analyzed directly by reading it and loading in the storage mechanisms introduce above in main memory. Therefore, copying a file to the DS server or workstation is all that is required. On the other hand, this phase represents a bottleneck in a database system and a more reasonable compromise in Hadoop systems. In a Hadoop system it can range from copying text files to formatting records in a specific storage format (key-value, Hive, Parquet). In a database system input records are generally encoded into a specific binary format, which allows efficient blocked random access and record indexing. A key aspect in a parallel system, is partitioning the input data set and distributing it, explained below.

3.3 Integrating Diverse Sources

Data science languages (Python, R) provide significant freedom to integrate data sets as needed. They have libraries to merge data sets by a common columns(s), behaving in a similar manner to a relational join. Python is more flexible than R to manipulate text (words, sentences). R provides better defined operators and functions to manipulate matrices. In the absence of columns with overlapping content, data integration is difficult. In a Hadoop system there exist libraries that attempt to match records by content, which is especially challenging with strings. The most principled approaches come from the database world, where records can be matched by columns or by content, but they are generally restricted to tabular data (i.e. relational tables).

3.4 Maintaining Data integrity: ACID

In data science languages, like Python and R, there is a vague notion of ACID properties, which tends to be ignored, especially with text data (documents, web pages). When there are significant data additions or changes, the data preparation pipeline is re-executed and the analytical tasks repeated with a refreshed data set. Hadoop systems have been gradually strengthened with more ACID properties, especially to explore data sets interactively. It is noteworthy ACID properties have gradually subsumed eventual consistency. DBMSs provide the strongest ACID guarantees, but they are generally considered a second alternative after Hadoop big data systems to analyze large volumes of data with machine learning or graph algorithms. DBMSs forte is query processing. In general, most DBMSs propagate changes on queries recomputing materialized views. As there is more interest in getting near real-time (active) results, but maintaining consistency, the underlying tables are locked or maintained with MVCC.

4 Processing

4.1 A Flexible and Comprehensive Processing Architecture

It is difficult to synthesize all different DS pipelines into a single architecture. Here we list the most common elements: diverse data sources, a data repository, an analytic computer or cluster. The tasks: loading, cleaning, integrating, analyzing, deploying. Data sources include: databases, logs, documents, perhaps with inconsistent and dirty content. The data repository can be a folder in a shared server or a data lake in parallel cluster. Cleaning and integrating big data is generally done combining Hadoop and DBMSs: it just depends where the bulk of information is. Some years ago the standard was to process big data in a local parallel cluster. Cloud computing [6] has changed the landscape, having data lakes in the cloud, which enables analysts to integrate and clean data in the cloud. In general a data set for analysis is much smaller than raw big data, which allow analysts to process data locally, in their own workstation or server.

4.2 Serial Processing

This is the norm in data science languiages. It is easier to develop small programs without worrying about low-level details such as multicore CPUs, threads and network communication. In general, each analyst runs in their own space, without worrying about shared memory. In a large project, data integration and preparation is assumed to have been take care of before.

4.3 Parallel Processing

This is the turf of Hadoop and parallel DBMSs, which compete with each other [4]. In most systems parallel processign requires three phases: (1) partitioniong

data and distributing chunks/blocks; (2) running processing code on each partition; (3) gathering or assembling partial results. Phase 1 may require sorting, automatically or manually and hashing records by a key. In general, the goal is to run independently on each partition and minimize coordination. There is a debate between scaleup (more cores, more RAM) and scale out (more machines, less RAM on each). In DS languages there is automatic parallelism in the object code which is optimized for modern multicore CPUs. Multicore CPUs, larger RAM and SSDs favor doing analysis in one beefy machine. On the other hand, big data, especially "Volume" support distributed processing to overcome the I/O bottleneck. Streams, being sequential data, lead to distributed filtering and summarization, but commonly centralized processing of complex models. Hadoop and DBMSs offer a parallel version of scan, sort and merge/join. However, Hadoop systems trail DBMSs on indexing and advanced join processing, but the gap keeps shrinking. When there are indexes, either in main memory or secondary storage filtering, joining and aggregations can be accelerated.

4.4 Fault Tolerance

There are two main scenarios: fault tolerance to avoid data loss and fault tolerance during processing. This is a contrasting feature with data science languages, which do not provide fault tolerance either to avoid data loss or to avoid redoing work when there are runtime errors. In many cases, this is not seen as a major limitation because analytic data sets can be recreated from source data and because code bugs can be fixed. But they represent a time loss. On the other hand, Hadoop systems provide run-time fault tolerance during processing when one node or machine fail, with automatic data copy/recovery from a backup node or disk. This feature is at the heart of the Google file system (proprietary) and HDFS (open source). Parallel DBMSs provide fault tolerance to avoid losing data, going from secondary copies of each data block to the internal log, where all update operations are recorded, It is generally considered that continuously appending the recovery log is required for transactions, but an overhead for analytics.

5 Analytics

5.1 Definitions

We introduce definitions for the main mathematical objects used in data science. We use matrix, graph and relational table as the main and most general mathematical objects. Graphs can represent cubes and itemsets as a particular case.

The most important one is the matrix, which may be square or rectangular. A matrix is composed of a list of vectors.

A tuple is an n-ary list of values of possibly diverse types, accessed by attribute name. A specific tuple in a relational table is accessed by a key. Finally,

bags and sets contain any kind of object without any pre-defined structure, where elements are unioque in sets and there may be repetitions in bags.

It is important to emphasize these mathematical objects are different. A matrix is different from a relational table. In a matrix elements are accessed by subscript, whereas in a relational table by key and attribute name. In an analagous manner, a vector is different from a tuple. Relational tables are sets of uniform objects (tuples), but sets in general are not. Then bags represent generic containers for anything, including words, files, iamges and so on. We will argue different systems tend to be better for one kind of object, requring significant effort the other two.

A graph $G = (V, E)$ consists of a set of vertices V and a set of edges E connecting them. Graph size is given by $n = |V|$ and $m = |E|$, where $m = O(n)$ for sparse graphs and $m = O(n^2)$ for dense graphs, with a close correspondence to sparse and dense matrices. A graph can be manipulated as a matrix or as a list of edges (or adjacent vertices).

5.2 Exploration

A data set is commonly explored with descriptive statistics and histograms. These statistical mechanisms and techniques give an idea about data distribution. When a hypothesis comes up it is common to run some statistical test whose goal is to reject or accept a user-defined hypothesis. In general, these analyses are fast and they scale reasonably well on large data sets in a data science language, especially when the data sets fits in RAM. Together with statistics Data Science systems provide visualization aids with plots, mesh grids and histograms, both in 2D and 3D. When interacting with a DBMS, analysts explore the data set with queries, or tool that generate queries. In general queries are written in SQL and they combine filters, joins and aggregations. In cube processing and exploration, one query leads to another more focused query. However, the analyst commonly exports slightly pre-processed data sets outside the DBMS due to the ease of use of DS languages (leaving performance as a secondary aspect). Lately, data science languages provide almost equivalent routines to explore the data set with equivalent mechanisms to queries. Out of many systems out there, the Pandas library represents a primitve, but powerful, query mechanism whose flexibility is better than SQL, but lacking important features of query processing.

5.3 Graphs

The most complex exploration mechanism is graphs, which are flexible to represent any set of interconnected objects. Nowadays, they are particularly useful to represent social networks and interconnected computers and devices on the Internet. Their generality allows answering many exploratory questions, which are practically impossible to get with descriptive statistics. Even though graphs represent a mathematical model they can be considered descriptive models rather than predictive. Nevertheless, many algorithms on graphs are computationally challenging, with many of them involving NP-complete problems or exponential

time and space complexity. Well-known problems include paths, reachability, centrality, diameter and clique detection, most of which remain open with big data. Graph engines (Spark GraphX, Neo4j) lead, followed by parallel DBMSs (Vertica, Teradata, Tigergraph) to analyze large graphs. DS language libraries do not scale well with large graphs, especially when they do fit in main memory.

5.4 Mathematical Models

Machine learning and statistic are the most prominent mathematical models. We broadly classify models as descriptive and predictive. Descriptive (unsupervised) models are a generalization of descriptive statistics, like the mean and variance, in one dimension. For predictive (supervised) models there is an output variable, which makes the learning problem significantly harder. If the variable is continueous we commonly refer to it as Y, whereas if it a discrete variable we call it G. Nowadays machine learning has subsumed statistics to build predictive models This is in good part due to deep neural networks, which can work on unlabeled data, on text and on images. Moreover, deep neural networks have the capability of automatically deriving features (variables), simplifying data pre-processing. Generally speaking these computation involve iterative algorithms involving matrix computations. The time complexity goes from $O(dn)$ to $O(2^d n)$ and $O(dn^2)$ in practice. But it can go up to $O(2^d n)$ for variable/feature selection or Bayesian variable networks. The number of iterations is generally considered an orthogonal aspect to $O()$, being a challenge for clustering algorithms like K-means and EM. Data science language libraries excel in the variety of models they and the ability to stack them. Models range from simple predictive models like NB and PCA to deep neural networks. It is fair to say Python dominates the landscape in neural networks (Tensorflow, Keras Scikit-learn) and R in advanced statistical models. Hadoop systems offer specialized libraries to compute models like Spark MLlib and Mahout. The trend has been to build wrappers in a DS language, where Python is now the dominant DS language Spark, leaving Scala for expert users. DBMSs offer some algorithms programmed via SQL queries and UDFs, fast cursor interfaces, or internal conversion from relational to matrix format, but they are difficult to use, and they cannot be easily combined and they require importing external data.

Given DBMS rigid archicture, and learning curve, users tend to prefer DS languages, followed by Hadoop with data volume forces the analyst to use such tool. That is, DBMSs have lost ground as an analytical platform, being used mainly for some strategic queries and in a few cases machine learning. In practice, many models are built on samples or highly pre-processed data sets where d and n have been reduced to manageable size. Building predictive models on big data is still an open problem, but it is a moving target towards DS languages.

6 Conclusions

Python and R are now the standard programming languages in Data Science. Clearly, Python and R are not the fastest languages, nor the ones that guarantee

ACID, but they are becoming more popular and their libraries faster. Among the two Python user base is growing faster, but R's vast statistical libraries (CRAN) will maintain its relevance for a long time. However, SQL remains necessary to query databases and extract and pre-process data coming from databases. Transactions remain an important source of new data, but non-transactional data is growing faster. Data warehousing is now an old concept, which has been substituted by so-called data lakes. Cube processing and ad-hoc queries are still used in specialized data warehouses, but they have been subsumed by exploratory statistical analysis in big data. Hadoop big data systems remain relevant for large volume and the cloud. Both Hadoop systems and DBMSs interoperate with Python and R, at different levels depending on the specific analytic system. It is fair to say CSV text files and JSON are now standad formats to exchange big data, leaving behind proprietary formats.

Data Science is opening many possibilities for future research, adapting and extending the state of the art in database systems, programming languages, data mining and parallel computing. Analysts require tools that are reasonably fast, intuitive and flexible. There should be seamless mechanisms to convert data formats, and avoid it when possible. Many research and tool developers are moving beyond the "fastest system" mentality, given hardware advances and the cloud, and they are focusing instead on reducing development time for the analyst. We need analytic algorithms and techniques that can interoperate and exchange data easily with relaxed structure assumptions.

References

1. H. Garcia-Molina, J.D. Ullman, and J. Widom. *Database Systems: The Complete Book*. Prentice Hall, 2nd edition, 2008.
2. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2006.
3. C. Ordonez and J. García-García. Managing big data analytics workflows with a database system. In *IEEE/ACM CCGrid*, pages 649–655, 2016.
4. M. Stonebraker, D. Abadi, D.J. DeWitt, S. Madden, E. Paulson, A. Pavlo, and A. Rasin. MapReduce and parallel DBMSs: friends or foes? *Commun. ACM*, 53(1):64–71, 2010.
5. M. Stonebraker, P. Brown, D. Zhang, and J. Becla. SciDB: A Database Management System for Applications with Complex Analytics. *Computing in Science and Engineering*, 15(3):54–62, 2013.
6. Y. Zhang, C. Ordonez, and L. Johnsson. A cloud system for machine learning exploiting a parallel array DBMS. In *Proc. DEXA Workshops (BDMICS)*, pages 22–26, 2017.