Jacobs University

Big Data Project

Big Data Age

Rahul Bhat

supervised by

Prof. Dr. Michael Kohlhase

February 21, 2016

Contents

Abstract

Demand for "Big Data" expertise started growing over the last few years and so as the demands for the new tools and the techniques that transform this huge amount of data into useful knowledge. The data volume is growing at a rapid rate in every industry or organization. So as "Big Data" related jobs are increasing so rapidly that every sector whether it is Finance, Engineering, Government, Consulting, Scientific, they all are looking for employees who are experts in Data Analysis and can handle "Big Data". The tons of data which are generated in these sectors could be structured or unstructured, and they need professionals who can work on this large quantity of data and turn that data into something valuable and the most crucial of these are "Data Scientists". Data Science is gaining traction in every sector and there are different profiles in Data Science Industry and everyone is an expert in its own field. The people who manage these profiles have a background in Computer Science or IT because Computer Engineering skills are needed to manage such a large volume of data. This research has been conducted on education needs, roles and responsibilities, skills, certifications and what type of Companies are hiring them. [12].

Keywords: Big Data, Data Science, R, Python, SQL

1 Introduction

In recent times, "Big Data" has been the talk in the analytic sector. The questions which one should ask is what is "Big Data"? Why is it so important? Where does "Big Data" come from? What's the relevance of "Big Data"? In this research paper, we have given a short summary of "Evaluation of "Big Data Technology" what in reality "Big Data" is and who handles "Big Data". There are experts who are managing and working on "Big Data". Their roles and responsibilities and the platforms on which they work are extremely different from each other. The Industries and Companies who are hiring these experts have different criteria and requirements. In this paper, we have explained in detail about the professionals who are working on "Big Data", their education background, the platform on which they are working on and, the type of Companies hiring them and their requirements. There are more than hundreds of tools available on market, in addition, we have tried to find out the popular and useful tools and the comparison between the best tools available on market.

2 Big Data Technology

2.1 History of Big Data

The word “Big Data” first mentioned in a book by Weiss and Indrukya in 1998 and the first academic paper mentioning ”Big Data” was appeared in 2000 by Diebold. The word ”Big Data” came because of the huge amount of data is generating every second. To understand the phenomenon that is ”Big Data” in 2000s industry analyst Doug Laney introduced the three Vs of data management, defining the three main components of data as volume, velocity, and variety. Volume refers to the vast amount of data generated per second. Variety refers to the different types of data we can use. Velocity refers to the speed at which new data is generated and the speed at which data move. The amount of data produced dramatically increases with time. As the years have continued IBM introduced the fourth V that is Veracity. Veracity refers to the biases, noise, and abnormality in data which lead to another challenge, keeping ”Big Data” organized. The data that being generated today leads to the fifth V and that is Value. The large amount of data is transformed into some valuable information and that information can be put to good use and yield value and business opportunities which could accelerate profitable growth [9].

2.2 What is Big Data?

Big Data is a massive volume of both structured or unstructured data, which moves so quickly and which exceeds the processing capacity of database systems. So it is very difficult to process this large amount of data using traditional techniques. The amount of data is growing so rapidly and the reason behind this is the fast-growing Internet speed, easy access to the Internet and advances in mobile devices that include digital video, photography, audio, and advanced email and text features and this data is generated by us, that is, 70% of the data is user-generated and the third reason is, Organizations and the Companies which are generating data at a fast speed. The digital universe is growing at a fast speed, in 2009, it grew to 60% that is almost 800,000 petabytes, In 2010, it was estimated 1.2 million and, in the end, it was in petabytes and, In 2020, it is predicted that the digital world will grow 44 times as big as it was in 2009 citecraig2011privacy. There was a time when data was used to count in kilobytes, megabytes, gigabytes, but nowadays data is estimated in zettabytes citejadhavdata. ”In 2005, the World Wide Web had 11.5 billion indexable documents. By 2007, that number grew to 25 billion, and by 2012 it doubled again, topping out at more than 56 billion” citedavissongoogle. The digital world is growing at such a speed that it became impossible for someone to predict the actual numbers and there would be time soon when these numbers would be countless.

In other words, one can say that ”Big Data” is an act of gathering and storing a large amount of data, but it is not only collecting and storing the data or arranging it in order, the important thing is after gathering and storing what benefits, an Organization will get from that data which would be fruitful. This large amount of data which has been collected from the different sources is then analyzed which gives the answers like smart decision making, cost reductions, time reductions, new product development and optimized offerings. This large amount of data is analyzed which would be helpful in determining the cause of failure, issues, and defects, what are the behaviors that affect the Organization and many more [14].

2.3 Who Handles Big Data?

The three important skills that are needed to be effective in handling "Big Data". First is an understanding of databases and how to manage large amounts of data. Next is knowledge about machine learning and data mining. Last comes statistics, so you can estimate the reliability of your conclusions. The question which comes in the mind is, who are these people who handle this "Big Data"? Every Industry big or small, have a large amount of data and they are experiencing a rush in the amount of data. The Companies and Organizations around the world need experts who can handle this large pool of data. So the organizations are hiring Data Engineers, who handles and work on this large amount of data.

2.4 Who is Data Engineer?

Data Engineers are the one who knows how to fish out the answers related to Business questions while swimming in the large pool of data. These are the people who make the new discoveries in the field of Data Science. A long time ago there were no University programs or courses for Data Science. The question was from where did the Organizations going to find these people with technical and analytic skills? The significant and growing demand for data professionals and rapidly rising salaries, Universities has recently begun to respond to the needs of Data Science program. Due to the increasing growth in Data Science, now more Universities are offering programs and certificates at Bachelors as well as at Master's level [6].

3 Discrete Roles and Responsibilities

3.1 Data Scientist vs Data Analyst

There are two main job profiles in Data Science one is Data Scientist and the second is Data Analyst. These twos are different fields and their roles and responsibilities are totally different from each other. They both use different tools and they work on the different environment. People and even some Organizations are confused between Data Scientists and Data Analysts. Some Organizations treat them like they both same or just the synonyms, but in reality, they both have their own terms of skill sets and experience. Organizations mixing these two different streams and confuse the job roles. However, these roles tend to be complementary to one another, but often span a wide variety of different skill sets and functional roles.

Somehow they both have the same mission in an Industry and even they both work on the large volume of data, but the thing is, they both use different techniques and tools to work on them. Data Analyst gathers the data, clean that data and then arrange that data in order and in proper structure (sort that data in a proper manner) and then that data has been forwarded to Data Scientist who uses the statistical skills and extract the valuable information from that data. For example, the Organization gives the task of Data Analyst to Data Scientist and the result is that they waste their time and effort in organizing, cleaning and sorting of a large amount of data.

3.2 Job Profiles

The other Data Science jobs which are different from each other are Data Architects, The Database Administrators, and Statisticians. They all have one thing in common and that is, they all are related to "Big Data" but the platform on which they work is extremely different from each other. A Data Engineer can also start his career with one of these Jobs and they are the strong candidates for these profiles.

3.3 Various Roles in Data Science

- Data Scientist
- Data Architect
- Statistician
- The Database Administrator

4 Primary Research

Data Scientist as a primary research has been conducted with the employee "Vijay Singh", who is working as Data Scientist at Verisk Analytics, New York. He is a graduate in Management Information Systems (MIS), but he has done his specialization in data mining and analytic. He has been extensively working in the field of data mining and analytic throughout the academic curriculum for his projects and publications as well as for a financial firm in the United States. With an education in technical and business domain. He has expertise in analytic and problem-solving skills. Before starting working as a Data Scientist, he did three certifications from SAS, which he said is very important and useful for Data Scientists. The details of these certifications are mentioned in this section.

4.1 Data Scientist

"Data Scientists" been called as the sexiest job of the 21st Century according to the Harvard Business Review Magazine" by Thomas H. Davenport and D.J. Patil. This is because data have swept into every Industry and Organization and the demand for Data Scientists are growing at a rapid rate. There is a high demand for Data Scientists in every sector whether it is Government or Private. Every Organization is looking for Data Scientists who can dig their data and extract the valuable information from that, which would be beneficial for their Industry. Data Scientists are very good in Data Visualization as well. Data visualization refers to representing complex data in a visual context, like a chart or a map, to help people understand the significance of that data. So "Data Scientists" know how to create charts, graphs and how to use the visualization tools. Machine learning is probably at the top of the required skills of Data Scientists. Data Scientists are very good thinkers which go beneath the surface of the material to finding the solutions. The "Data Scientists" are the "Data Artist" in terms of knowing how to choose the best way to visualize and present the discovered patterns and data associations. Data Scientists use the technical and analytic skills and extract meaning from a large pool of data. They are able to bring structure to large quantities of unstructured or unsorted data and make analysis possible [6].

4.2 Academic Background

Data Scientists are the Computer Engineers who have solid background typically in modeling, machine learning, statistics, analytic, and maths. Data Scientist can also emerge from the field of data and computation focus. Coding is the most basic universal skill of Data Scientist. Data Scientists are good statisticians and mathematician. They also have knowledge or experience in programming languages and database technologies such as Python, R and MySQL.

4.3 Roles and Responsibilities

The job responsibilities of a Data Scientist depends on the sector to sector and even from a company to company, in general though, the Data Scientist's role is to identify rich data sources, join them with the data sources which are incomplete and clean the resulting set. Data Scientists are half analysts and half artists who stares at the data, analyze that data from many angles and come up with a good solution to a business problem which would be beneficial for an Organization.

Today the job of Data Scientist is not only collecting and reporting on data and answer the questions and exploring existing assumptions and processes, but also be able to communicate their findings and recommendations in such a way that the Organization or Industry can understand and act on. So, there are varied roles and responsibilities of Data Scientists and those are :

- Data Scientists develop and plan analytic projects, designs experiments, conduct researches, test hypotheses and build models.
- Data Scientists develop new tools and analytical methods.
- Data Scientists conduct researches and moderately complex designs algorithms.
- Data Scientists works with the stakeholders to identify the business requirements and the expected outcome.
- Data Scientists take care of the needs of the client by working alongside with the business analysts.
- Respond and resolve the issues related to data mining and also provides quality assurance for data mining.
- Data Scientists also do the data visualization and comes with the proper presentation which is helpful for the people to understand the significance of the data.
- Data Scientists contribute to the data mining architectures the methodologies.
- Develop and plan required analytic projects in response to business needs.
- Data Scientists also works with a product engineering team to identify and to answer important product questions.

4.4 Tools Required

In "Data Scientist" survey, it has been found out that there are two tools which are most important that are "R" and "Python". "R" is the most popular tool among data miners, although Python usage is also

growing rapidly. There is also a big increase in Hadoop tool usage and there is the fastest growth for Spark as well. There are many other tools which are useful like SQL, Pandas, Julia, Tableau, Excel and many more. The Tools which are useful and most popular in Data Science field are as follows :

R - R is developed at Bell Laboratories (formerly AT&T, now Lucent Technologies). R is a language, which is used for graphics and statistical computing like linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering and many more. R can be easily extended and there are about eight packages in R and many more are available through CRAN family of Internet sites covering a very wide range of modern statistics. R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, One of the R's strength is the ease to use mathematical symbols and formulae. R is available as free software for both Windows and Mac-OS [7].

Python - "Python was conceived in the late 1980s and its implementation was started in December 1989 by Guido van Rossum". Python is an object-oriented, high-level programming language with dynamic semantics. Python incorporates modules, high-level built in data structures, dynamic typing, dynamic binding, exceptions, high-level data types and classes which make it very attractive and the plus point of this programming language is, it is easy to learn and portable. The Python interpreter and the extensive standard library are available in source or binary form for all major platforms for free which can be freely distributed. It is so user-friendly that it is not only using in academic settings but also in the Industries or Organizations[15].

MapReduce - More and more information started generation on the web and it was becoming very difficult to index over one billion pages of content and data processing became difficult, in order to cope, Google invented a new style of data processing known as MapReduce. A MapReduce computation executes as follows.
The computation takes a set of input key-value pairs from distributed system and produces sequences of output key-value pairs. MapReduce performs two tasks one is Map task and other is Reduce task.
The Map is written by the user, which takes an input from the distributed system and produces a set of key-value pairs. There is a master controller which collects these set of key values and passes them to the Reduce function.
The Reduce function is also written by the user, which accepts these set of key-value pairs. After accepting, it merges these values together to form a possibly smaller set of values. Typically just zero or one output value is produced per Reduce invocation. These set of smaller values is then supplied to the user's reduce function via an iterator. This allows a user to handle lists of values that are too large to fit in memory [10].

Hadoop - Hadoop is an open source project hosted by Apache Software Foundation. This software stack introduces entirely new economics for storing and processing data at scale. Hadoop provides massive storage for any kind of data, enormous processing power and has the ability to handle limitless concurrent tasks or jobs. When Google published a white paper describing the MapReduce framework, a year after that, Doug Cutting and Mike Cafarella inspired by the white paper and they both created Hadoop. They put the concept of MapReduce to an open-source web search engine called Nutch which handles multiple tasks simultaneously. Hadoop divides the data set and then run it in parallel over multiple nodes. Hadoop distributes its computation which helps in solving the problem of big data set which doesn't fit in a single computer [17].

Julia - Julia is an alternative to R, but completely different from that. Julia is a relatively new programming language for scientific computing. Julia is a high-level, high-performance dynamic programming lan-

guage for technical computing. Julia is garbage-collected and it provides a sophisticated compiler, distributed parallel execution, numerical accuracy, and an extensive mathematical function library. Julia's Base library is mostly written in Julia itself and it allows concurrent, parallel and distributed computing. It also integrates C and FORTRAN libraries for linear algebra, random number generation, signal processing, and string processing. It is an open source project, so all the codes are available on GitHub [2].

MySQL - Microsoft SQL is an open source relation database system (RDBMS) with fairly high stability, accessible support, reliable, security, high-performance and cost effective. The main advantage of MySQL is that it is user-friendly and super good at latency for running ad hoc queries. It is the world's most widely used RDBMS, and the most widely used open-source client–server model RDBMS. MySQL runs on many platforms and is stable. The documentation for MySQL is excellent. MySQL is popular for use in web applications because MySQL is fast and can easily work with Internet speed. MySQL is also good in other tasks especially in Enterprise-level applications, Open-source support, Low overhead and Available large table size. So, MySQL is an open source RDBMS that offers many advantages over other RDBMS with few disadvantages [13].

Shark - Shark is also known as *C#* and pronounced as "see-sharp". *C#* is an objected oriented and type-safe programming language. It is somehow same as C++ but it is different than Java. The design of *C#* is closer to C++ because most of its operators, keywords, and statements of *C#* is directly from C++. One of the differences between *C#* and C++ is that in *C#* the enums are type-safe and enums are not just integers. *C#* language is component-oriented, which means a user can add the concepts itself at the time of writing components in the language. Concepts such as properties, methods, events, attributes, and documentation are all first-class language constructs. *C#* is the first language to incorporate XML comment tags that can be used by the compiler to generate readable documentation directly from source code. There are many important features of *C#*. The one of the important features of *C#* is that "CLR (Common Language Runtime) is the virtual machine component of Microsoft's.NET framework, which manages the execution of .NET programs. A process known as just-in-time compilation (JIT) converts compiled code into machine instructions then these machine instructions are then executed by the CPU which are used by *C#* applications which allow multiple programs to execute in a single hardware address space. In *C#* the user can add concurrency to the program in order to reduce the latency of operations [4].

Tools used for Visualization

- Matplotlib - It is used for plotting in python.
- ggplot2 - It is also used for R.
- Tableau - It is good, simple and easy to use.
- QlikView - It is a separate physical server used to create test reports and dashboards before going into production.

The tools which we have discussed above are the common and the important tools. Additional tools listed below are not as popular than the above-mentioned, but very useful for Data Analysis.

SAS 9.4, SAS Enterprise Guide, SAS Enterprise Miner, SAS Forecast Studio, IBM SPSS Modeler, JMP, Apache Giraph, Pig, HBase and Pandas

4.5 Technical Skills Required

Data Mining Techniques, Predictive modeling, Statistical methods, Statistical data analysis, Regression model, Probability, Decision trees, Classification, Segmentation, Cluster analysis, Time series analysis, Text mining, Sentiment analysis, Association analysis, Link analysis.

4.6 Certifications Required

- SAS Certified Base Programmer Certificate
- SAS Certified Business Analyst Certificate
- SAS Certified Predictive Modeler using SAS Enterprise Miner

"SAS Institute is an American developer of analytic software-based in Cary, North Carolina. SAS helps suer to access, manage, analyze and report on data to aid in decision making". SAS (Statistical Analysis System) is a Data Science certification program for those, who wants to gain deep knowledge that is necessary to work as a Data Scientist. The SAS certified professional program offers five globally recognized certification. There are two SAS programming certifications (Base and Advanced) and three on Web Development and Warehouse Technology. To get the certifications, the candidate has to appear for the examination and after completing that exam, the candidate can be fully certified. To get this certification, the one should have at least a Master's degree or higher in a quantitative or technical field. This certification is very helpful and important for the person who wants to work as a Data Scientist. After getting the certification, the probability of getting the job as a Data Scientist increases because this certification is globally recognized. The certifications are not easy, each and individual certifications has different topics [16].

To get the base certification, the base exam has 25 subtopics are given within these five categories:

- Accessing Data
- Creating Data Structures
- Managing Data
- Generating Reports
- Handling Errors

To get the advanced certification, the advanced exam lists 18 separate topics under these three headings:

- Accessing Data Using SQL
- Macro Processing
- Advanced Programming Techniques

SAS Certification can play an important role in developing the career of Data Engineer. These certifications are not only important for Data Scientists but also important for other profiles. The topics which are mentioned above are the most important skills and requirements required in a Data Science Industry.

4.7 Job Requirement

- Strong written and verbal communication skills.
- Degree in a quantitative field with working knowledge of CS.
- Knowledge of RStudio, Python, SAS and/or R, Hadoop, Python, MySQL.
- Having the ability to query databases and perform statistical analysis.
- Being able to develop or program databases.
- Having an, at least basic, understanding of how a business and strategy works.
- Being able to create examples, prototypes, demonstrations to help management better understand the work.
- Having a good understanding of design and architecture principles.
- Being able to work autonomously.

4.8 Types of Companies Hiring Data Scientists

- Government Agencies
- Technology Companies
- Consulting and Research Firms
- Scientific Organizations

Data scientists are highly educated – 88% has at least a Master's degree and 46% has Ph.D's. Every company will value skills and tools a bit differently, but they are looking for a person who has a good knowledge and experience in Python, R and MySQL. If one has experience in these three areas, then he will be making a strong case for himself as a Data Science candidate. The knowledge of these three languages are required, to get a job as a Data Scientist. In short, the Big Data Scientist needs to have an understanding of almost everything and it also depends on in which field or industry the Data Scientist wants to work.

5 Data Warehousing and Data Modeling

5.1 Data Architect

In an Industry or an Organization, we are familiar with so many titles or profiles related to Architect. There are Software Architects, Infrastructure Architects, Application Architects, Business Intelligence Architects, Data Architects, Information Architects, and more. When it comes to the word Architect one thing comes to our mind that is design or structure. So Data Architect is an important role in an Organization, who is responsible for the design, structure and maintenance of data. Data Architects works closely with the clients and the developers by evaluating and implementing their needs and suggestions. Data Architects are the one who has the knowledge of Data Warehousing and are experts in Data Modeling techniques. [8].

Data Warehousing - Data warehouse is the architecture designed for information systems. A Data Warehouse is actually a relational database that is designed for query and analysis. In simple words, operational system is where we put the data in, and Data Warehouse is from where we get the data. So Data Warehouse is a repository, which contains historical data derived from transaction data, but it also includes the integrated data from multiple heterogeneous sources. It separates analysis workload from transaction workload and it also organizes and stores the data for informational and analytic processing. "Data Architects" gather data, analyze it and take decisions based on the information present in the Data Warehouse. The data stored in the Data Warehouse can be used for the domains like tuning production strategies, customer analysis, and operations analysis. Data Warehousing also involves data cleaning, data integration, data consolidation, data transformation, data loading and refreshing [11].

Data Modeling Technique - The title "Data modeling" itself describes that it is something related to modeling of data. A model is nothing but the reflection of the real world and modeling gives us the ability to visualize what we cannot realize. It is the same with Data Modeling. Data modeling is concerned with describing, organizing and analyzing the data. Data Modeling organizes the enterprise data for communication between developers and describing the data in a data model by representing it in a diagram. It is the act of exploring data-oriented structures. Data Modeling helps the user to identify the entity types like class modeling and to identify class types. There are some tools which are useful in Data Modeling which makes it easier to design a database by capturing business data and their relationships and enforce data integrity with business rules [8].

ER Diagrams - "Data Modelers" use ER Diagram (ERD) the most, for the data modeling process, as a communication media with the business end users. "ER" stands for "Entity" and "Relationship", and it shows the relationships of entity sets stored in a database, where "Entity" is the component of data. ER Diagram looks the same as a flowchart, but with specialized symbols, where symbols have some meaning. ER Diagram helps Data Modelers in designing the resulting data structure. ER Diagram helps end users to easily understand and navigate the data structure and fully exploit the data [18].

5.2 Academic Background

Data Architects are experts at both defining data structures and at data integration practices. Data Architects could be Bachelor's and it is not compulsory that the minimum qualification should be at least Master's to become a Data Architect like Data Scientists, but the important thing is, they should be from Engineering background with major in Information Technology or Computer Science. They need to be technically strong and have good knowledge of Computer Programming with some knowledge in business studies as well.

5.3 Roles and Responsibilities

- Multinational companies have two type of sources one is Internal sources and other is External Sources. External such as market feeds and Internal such as existing systems. Data Architects designs a plan to maintain, integrate and build conceptual, logical and physical database models on these (internal and external) data sources.
- Data Architects gives the representation of the database models.
- Data Architect should know where this data is coming from and how to structure that data in such

a way that it could be useful for an Organization.

- Data Architects manage outsourcing vendors, review designs, code and provide architecture leadership and guidance.
- Data Architects defines and implements security layers in all warehouses and ensure new code and architecture.
- Data Architects built and designed a monitoring database to capture middle tier usage statistics.
- Data Architects work closely on the Enterprise Data Model and provide input for corporate Data Architecture standards.
- Data Architects design encryption process to test and development data.
- Data Architects work with the sales team to create and deliver client proposals and demonstrations.
- Data Architects mentor development team members in design and development of complex ETL implementations.
- Data Architects assigns resources to projects, provides technical guidance and handle personnel issues.
- Data Architects ensure all warehouse projects are adhering to corporate naming, architecture and performance standards.
- Data Architects Implement database solutions using available database development tools.
- Data Architects has to provide a written status reports about the project status, tasks and if there are any issues and risks to the management.

5.4 Tools Required

There are more than 50 database modeling tools in the market, but there are some tools which are most popular and important for Data Modeling Technique. Below you find the favorite Data modeling Tools :

Java, Oracle, Hadoop, NoSQL, JSF, JQuery, OpenXava, JSON, SOAP, Tableau, R, SQL, ERwin, ER/Studio, PHP, Python, Matlab and good knowledge of ETL tools.

Today is the world of competition and in this increasingly heterogeneous environment, particularly when there are hundreds of Data Modeling tools available in the market. It is very difficult to find the best among them. It is very important that the tool should have some other more general characteristics like the tools should be flexible, compatible and have the ability to interchange with other tools and strong sets of documentation and online resources. So there are some tools with these capabilities and characteristics, and these tools are ER/Studio, ERwin, and ETL tools which are the most important and widely used tools by Data Architects. ER/Studio and ERwin, these two tools are used for Data Modeling Technique and ETL tools are used for Data Warehousing.

ER/Studio - ER/Studio is known as the award-winning data modeling application for logical, and physical database design and construction. ER/Studio is powerful and it has a multi-level design environment for Database Administrators, Developers, and Data Architects who build and maintain large, complex database applications. The main feature of ER/Studio is that the certifications, which build on ER/Studio

supports for a variety of database platforms including Apache Hive, Hadoop, MongoDB Enterprise, and traditional relational database management systems (RDBMS). There are so many features in ER/Studio such as indexing, transforms, and forward engineering to turn a logical data model into an efficient physical design. User are able to read a data model of any size and complexity with the same confidence as reading a book.

CA ERwin - CA ERwin also known as ERwin and it is a Data Modeling tool which provides a simple, visual interface to manage your complex data environment through a graphical interface. ERwin has a feature called Bulk Editor, which has a function that enables users to quickly and easily view and update with multiple object types, or multiple properties within a single object type. There are the two most important features provide by ERwin and that is Forward Engineering and Reverse engineering. Forward Engineering transforms the data model into database, and Reverse Engineering is obtaining a design information and data model from a relational database.

ETL Tool - ETL tools are very important and useful tools for a Data Architects. ETL tools are mostly used in Business Intelligence. ETL stands for, Extract, Transform and Load. It is a process of extracting the data from the source then transform that data into a proper format and then loading that transformed data into Data Warehouse.

The List of the most popular ETL tools are as follows :

- Oracle Warehouse Builder (OWB)
- SAP Data Services
- IBM Infosphere Information Server
- SAS Data Management
- PowerCenter Informatica
- Elixir Repertoire for Data ETL
- Data Migrator (IBI)
- SQL Server Integration Services (SSIS)
- Talend Studio for Data Integration
- Sagent Data Flow
- Pervasive Data Integrator
- Open Text Integration Center
- Oracle Data Integrator (ODI)
- Cognos Data Manager
- CloverETL
- Centerprise Data Integrator
- IBM Infosphere Warehouse Edition

- Pentaho Data Integration
- Adeptia Integration Server
- Syncsort DMX
- QlikView Expressor
- Relational Junction ETL Manager (Sesame Software)

So these are the Tools and Skills which Companies or Organizations mostly look into a Data Architect. The company's desire to have someone who has a very good knowledge of ETL tools.

5.5 Technical Skills Required

Data Analysis, Data Migration Tools Knowledge, Data Modeling, Data warehousing, Database Design, Data Integration, Business Rule Management, Business Intelligence Reporting.

5.6 Job Requirement

- Strong knowledge of IT systems, practices and functional units.
- Knowledge of Data Warehouse principles and methodologies.
- Knowledge and expertise of Database Modeling Techniques.
- Knowledge of best practices in the Database, ETL and/or BI Technologies.
- Knowledge of Entity Framework, SQL Server Data Tools/Business Intelligence, Analysis Services, Integration Services, Reporting Services, and Profiler.
- Provides technical consultation to the business units that they support.
- Experience with modeling tools like ER/Studio, ERwin.
- Knowledge and experience with BI reporting and visualization tools - Business Objects, Microstrategy, Cognos, Tableau, Qlikview, XLcubed, MSBI(SSAS, SSRS, Performance Point).
- Strong Oracle experience and skills including stored procedure design and development.
- Guiding and overseeing the database design process which includes researching and recommending optimal design criteria for both transaction and data warehousing models.
- Reviewing the overall physical database structures for data integrity, performance quality, recoverability, maintenance, and space requirement considerations.
- Developing rules, procedures, and standards for the access and maintenance of shared data resources.
- Eagerness to learn new tools and technologies.
- Experience building EIM or BI roadmaps to meet business needs.

- Working with IT and support to ensure that the database system is operating in accordance with required system usage.
- Excellent written and verbal communication, team building, and relationship building skills.
- Ability to manage and complete multiple technical deliverable on aggressive schedules.
- Strong communication skills.

Data Architect should be good in Data Warehousing and Data Modeling Techniques. Every industry or organization has their own requirements, some want to hire a Data Architect who is expert at Data Modeling Techniques and some wants to hire who is, skilled in Data Warehousing practices. It depends on the organization and their needs. The main job of a Data Architect is to change the Organization's data systems to mirror the model.

5.7 Types of Companies Hiring Data Architects

- Government Agencies
- Technology Companies
- Consulting and Research Firms
- Scientific Organizations

6 Data and Statistics

6.1 Statistician

The increasing demand of Big Data in engineering motivates the Department of Statistics. Statisticians are a diverse group of people with one thing in common, they use statistics to draw valuable insights from data. Statiscians are not good programmers but Data Scientists are not only good Statisticians, but they are good in writing software programs as well. The job of Statistician is to gather the data, arrange that data in proper manner and then apply linear regression on that data to produce trustworthy information from that. Statisticians know how to find correlations, compute different types of regression, and understanding probability distributions. Statisticians analyze the data, data are not just the numbers, but numbers carry meaningful information which could be an asset for an Industry. Statisticians work is to deal with the numbers and extract meaning from that. Statisticians after analyzing the data and getting meaningful information they create charts, tables and draws practical conclusions from that data. [3].

6.2 Academic Background

Statisticians are the one who has strong science background with major in statistics or applied mathematics. They could be Bachelor's, but Graduate Degree is recommended, it is not compulsory that the statistician should be from the Engineering background. Scientific knowledge as well as technical knowledge is very important to become a good Statistician, because scientific knowledge helps Statisticians to understand the subject matter and technical background helps, to solve the problems easily. Statisticians

understand statistics theoretically and then applied them to the real life. Statisticians can work in any department they are interested in like mathematics, bio-statistics, public health, psychology, engineering, education, business, and economics.

6.3 Roles and Responsibilities

- Statisticians apply statistical theories and methods to solve the practical problems of various industries.
- Statisticians decide which data would be helpful, which is needed to solve the particular problem.
- Statisticians apply methods, design surveys or experiments to collect data.
- Statisticians develop and initiate innovative statistical techniques, issues and protocols.
- Statisticians develop easy-to-analyze sampling techniques and processes.
- Statisticians execute statistical operations in total fairness to derive zero-error results.
- Statisticians lead, guide and mentor statistical assistants in their day-to-day tasks.
- Statisticians utilize software and appropriate tools to perform statistical analysis.
- Statisticians organize and analyze samples, data sets and models and other issues.
- Statisticians identify and determine the type and kind of studies and research.
- Statisticians analyze and perform statistics on abstracts, data sets and other related information.
- Statisticians ensure compliance of standards and procedures of business client.
- Statisticians integrate best practices in statistical performance.
- Statisticians consult with clients and agreeing what data to collect and how it should be gathered.
- Presenting results to senior managers, regulatory authorities, clients.
- Statisticians after gathering the data, rearrange it in proper way and then report conclusions from their analysis.

6.4 Tools Required

Rstudio, Python and Matlab.

6.5 Technical Skills Required

- Technical and Analytical.
- Ability to Analyze, Model and Interpret Data.
- Statistics, Calculus and Linear Algebra.
- Operational Research.

- Mathematical Ability and Computer Literacy.
- Understanding of Statistical Terms and Concepts.

Statisticians don't need enough programming languages, the only thing that is important, they should be good in Linear Algebra and Statistics with experience or good knowledge of Rstudio, Python and Matlab. So, Statisticians use logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions, or approaches to problems.

6.6 Job Requirements

- Good knowledge of Matlab, R and Python.
- Able to use specialized computer software to organize and analyze data.
- Must have an aptitude for and interest in mathematics.
- Must have an interest in the application of scientific principles to the solution of practical problems.
- Must be able to organize projects and carry them out.
- Written and oral communication skills.
- Must be able to write clear, concise reports in language appropriate for intended audience.
- Problem-solving skills.
- The ability to communicate results and findings to non-statisticians.
- The ability to influence others.
- Practical and strategic approach to work.
- The capability to work to deadlines and to plan your work.
- The ability to work alone and within teams.

6.7 Types of Companies Hiring Statisticians

- Financial institutions
- Statistics Canada and other government departments
- Medical research agencies
- Engineering consulting firms
- Environmental consulting firms
- Environmental conservation agencies
- Market research companies
- Pharmaceutical companies
- Companies requiring process improvement
- Colleges and universities

7 DDL and DML

7.1 The Database Administrator (DBA)

The Database Administrators are the one who performs all activities related to the Database in an Organization or Industry. Their minimum qualification is Bachelor's in Computer Science or IT because they need good programming skills and good technical background. The Database administrators are responsible for the Data Maintenance, Performance, Data Integrity, Security. They are also responsible for the planning, designing, implementing and development of a new Database application. "Database Administrators" ensure that the data is accurate, available and secure. The Database Administrator's Job profile is very important in an Organization. The Database Administrator also manage Organization's data and meta-data. There are three important things in which Database Administrator should be strong or experienced in and that is Structured Query Language (SQL), SAP and Oracle-Based Database Management Software.

7.2 Academic Background

Database Administrator requires Bachelor's Degree in Computer Studies or Software/Computer Engineering and for graduates without relevant qualifications or experience and for postgraduate Computer/IT qualification is beneficial. They don't need knowledge in statistics but they need good programming skills.

7.3 Roles and Responsibilities

- Creation and Maintenance of database for Production, Testing and Development.
- Installation of oracle database software and patches.
- Managing Oracle databases on Standalone and Data Guard environment.
- Setting up Oracle Data Guard for high availability of databases by creating physical standby databases, troubleshooting and managing them.
- Monitor data activities (i.e. database status, logs, space utilization).
- Create users, tables spaces, grant roles and privileges.
- Monitor database health, regular backups on daily basis.
- Performing upgrades of the database and software to new release levels.
- Planning, scheduling, monitoring and troubleshooting backup and recovery procedures.
- Managing the databases on Automatic Storage Management (ASM).
- Database cloning (Manual and RMAN).
- Stats gathering, creation of indexes.
- Implement and monitor scheduled jobs.
- Schema, table-space management.

- Performed different recoveries, database flashback.
- Reorganization of tables and indexes.
- Monitoring database growth to ensure smooth functioning of database.
- Implementation of database monitoring scripts to get email notifications.

7.4 Tools Required

MySQL, SQL db, Database principles (ACID), DDL, DML, ETL Tools. The most important tools used by the Database Administrator are SQL, DDL, DML and ETL tools.

DDL and DML
SQL statements are divided into two major categories one is Data Definition Language (DDL) and the second is Data Manipulation Language (DML).

DDL - Data Definition Language (DDL) is a vocabulary used to define Data Structures in SQL Server. DDL is a subset of SQL statements that can change the whole structure of the database schema by creating, deleting or modifying schema objects such as tables, indexes by using the keywords CREATE, DROP or ALTER.
Examples are given below :

- CREATE - To create objects in the database.
- ALTER - Alters the structure of the database.
- DROP - Delete objects from the database.
- TRUNCATE - Remove all records from a table.
- COMMENT - Add comments to the data dictionary.
- RENAME - Rename an object.

DML - The DML (Data Manipulation Language) is also a vocabulary like DDL, but the difference is, it is used for managing data within schema objects. DML is also a subset of SQL Statements that can change the structure of the database schema by adding, removing, modifying by using keywords SELECT, INSERT, UPDATE OR DELETE.
Examples are given below :

- SELECT - Retrieve data from the a database.
- INSERT - Insert data into a table.
- UPDATE - Update existing data within a table.
- DELETE - Delete all records from a table.
- MERGE - UPSERT operation (insert or update).
- CALL - Call a PL/SQL or Java subprogram.
- EXPLAIN PLAN - Explain access path to data.
- LOCK TABLE - Control concurrency.

7.5 Technical Skills Required

- System Analysis and Design Skills.
- Database Design Skills.
- Physical Disk Storage Skills.
- Data Security Skills.
- Backup and Recovery Skills.
- Change Control Management Skills.
- Personal Time Management Skills.

7.6 Certifications Required

Oracle database administrator and MS (Microsoft) SQL Server. These are the two most important certification series for database professionals.

Oracle database administrator

- Oracle Database 12c: SQL Fundamentals 1Z0-061
- Oracle Database 11g: SQL Fundamentals I 1Z0-051
- Oracle Database SQL Expert 1Z0-047

Oracle Database 12c: SQL Fundamentals 1Z0-061

In this course students learn about Cloud Computing and this certification emphasize the full set of skills that DBA's need in today's competitive marketplace.

Objectives

- Retrieving data using the SQL SELECT statement.
- Restricting and sorting data.
- Using single-row functions to customize output.
- Reporting aggregated data using the group functions.
- Displaying data from multiple tables.
- Using subqueries to solve queries.
- Managing Tables using DML statements.
- Introduction to Data Definition Language.

Oracle Database 11g: SQL Fundamentals I 1Z0-051

In this course students learn the concepts of relational databases and the powerful SQL programming language like how to write the queries, how to manipulate data in tables and how to create database objects. This course also includes DDL and DML statements.

Objectives

- Retrieve row and column data from tables with the SELECT statement.
- Create reports of sorted and restricted data.
- Display data from multiple tables.
- Use DML statements to manage data.
- Use DDL statements to manage database objects.

Oracle Database SQL Expert 1Z0-047

This complete set of skills required for working with SQL programming language and have mastered the key concepts of a relational database.

Objectives

- Retrieving data using the SQL SELECT statement.
- Restricting and sorting data.
- Using single-row functions to customize output.
- Displaying data from multiple tables.
- Using subqueries to solve queries.
- Using the set operators.
- Manipulating data.
- Using DDL statements to create and manage tables.
- Creating other schema objects.
- Managing objects with data dictionary news.
- Controlling user access.
- Managing schema objectives.
- Manipulating large data sets.
- Managing data in different time zones.

MS (Microsoft) SQL Server

There are three Certification under SQL Server

- Microsoft Technology Associate (MTA) - Entry Level
- Microsoft Certified Solutions Associate (MCSA) - Associate level
- Microsoft Certified Solutions Expert (MCSE) - Expert level

Oracle has its own institution known as "Oracle University" and the institute of Microsoft is known as "Microsoft Virtual Academy" which are located all over the World, where they offer these courses.

7.7 Job Requirements

- Ability to understand and apply analytical and statistics tools.
- Experience in an Oracle Database 10G/11G/12C in a RAC, ASM and Data Guard environment or the Oracle Certifications in 10G/11G/12C.
- Good skills in all Oracle tools.
- A good knowledge of Oracle security management.
- A good knowledge of how Oracle acquires and manages resources.
- Excellent knowledge of Oracle backup and recovery scenarios.
- Strong verbal and written communication skills.
- Ability to assess and resolve complex technical issues.
- Ability to work independently.
- Effective facilitation, communication and teaching skills.
- Ability to engage and influence the organization.
- Ability to multi-task and handle large workloads under time constraints.
- Be able to provide a strategic database direction for the Organisation.
- Ability to perform both Oracle and also operating system performance.
- A good knowledge of physical database design.
- Ability to multi-task and handle large workloads under time constraints.

7.8 Types of Companies Hiring Statisticians

- Finance and Insurance
- Computer Systems Design
- Management of Companies and Enterprises
- Information Technology

- Educational Services: State, Local, and Private

The Database Administrator must know about Structured Query Language (SQL), SAP and Oracle-base Database Management Software. Oracle database administrator and MS (Microsoft) SQL Server are two Certifications which are very important for a Database Administrator. These two certifications will be an advantage for any Database Administrator for his career.

8 Advantages and Disadvantages of Big Data

Advantages - There is no doubt that there are amazing benefits from this growing digital world of "Big Data". Big Data is everywhere, from healthcare to sports, to the way we elect a president. "Big Data" has made big changes in our life. Big Data Technology has helped us understand and predict our collective behaviors such as illnesses can be identified or tracked before they worsen, now air pollution can be controlled which reduces energy consumption. Big Data has changed the behavior of every Organization or Industry by helping the Companies to create stronger connections with their customers and find leaner ways of operating.

The three main benefits of Big Data s are, faster and better decision making, new products and services and the third is cost saving. Big Data has enable Organizations to gather huge reams of information to help provide better insights and so make better decisions. Companies and Industries are promoting new products and services. Online companies have done this for a decade, but now predominantly offline firms are doing it too. Managing prodigious volumes of data is not only challenging for Organizations point of view, but it's often expensive as well. There are Real-Time Data Analytic tools available which Companies or Industries are implementing to eventually save a lot of money, time and also reduce the burden on Companies overall IT landscape. The most important and beneficial advantage is, fraud can be detected the moment it happens and better decision can be taken.

Online companies have done this for a decade or so, but now predominantly offline firms are doing it too. According to "Big Data Use Cases 2015, over 40% of Companies Worldwide analyze benefits from Big Data Technology. The Companies analyzed the list of benefits from "Big Data" analysis are better strategic decisions, improved control of operational processes, better understanding of customers and cost reductions. Furthermore, those organizations able to quantify there gains from analyzing Big Data reported an average 8% increase in revenues and a 10% reduction in cost [1].

Disadvantages - While the potential of "Big Data" is enormous, the main concern is over the Privacy and Security. Quintilian bytes of data is generating each day and it raises the issue of Security and privacy, which are magnified by velocity, volume, and the variety of "Big Data". "There are one trillion unique URLs in Google's index, two billion Google searches every day, 70 million videos available on YouTube, 133 million blogs, more than 29 billion tweets and more than 500 million active user's on Facebook who spends over 700 billion minutes per month on the site". The Institutions and Companies whether it is Government or Private, they store a large amount of data they are producing in a cloud. Information regarding individuals health, location, electricity use, and online activity is exposed, this use of large-scale cloud infrastructures which are spread across large networks of computers, with a diversity of software platforms, also increases the chance of Privacy and Security. Data helps in better decision making and to make accurate decisions, in time, it becomes necessary that it should be available in accurate, complete and timely manner [5].

9 Comparison Between Data Science Tools

R and Python

There's always been a battle between R and Python. Data Scientists often debate on the fact that who has better usability, features, and the most important is, about their libraries however, both the programming languages have their specialized key features complementing each other. They both have Pros and Cons and selecting over other depends on the use-case and cost of learning.
Following are the Pros and Cons :
R was developed for statisticians and can be experienced as slow due to poorly written code and there are some packages in R which need to be improving. Python is a general-purpose programming language with the easy syntax which increases the speed at which user write a program, on the other hand, R is a very simple and data focused in comparison, for example, R is not optimized for loops. Visualization is very important and Python has good visualization libraries like Seaborn, Bokeh and Pygal whereas, R is more flexible and R has the most advanced graphical capabilities among all languages. R and Python has the biggest online community, but there is no customer service support. So, if user have trouble regarding something then, there is no support from their side, although one can get help through online.

Overall, Python and R are among the popular programming languages that a Data Scientist must know to pursue a lucrative career in Data Science.

SQL and NoSQL

SQL is a universal language for managing and handling structured data and is famous for its ease of use, power and flexibility, which is beneficial at data collection and cleaning and transformation phases of the workflow. In SQL schema is required to define tables prior to use whereas, in NoSQL there is no schema required to store data. One of the important features of SQL is transactions because transactions ensure that you have atomically made changes to your database which is also very important whereas, NoSQL platforms don't support transactions. Scaling is the biggest issue in SQL because the relational database has to be on servers which are powerful and can handle relational database, but that server are expensive and difficult to handle. Scaling out is distributing databases across multiple hosts and handling tables across different servers is not an easy job, but NoSQL does this better because it is designed for optimal use on scaled out databases. NoSQL provides excellent performance and scalability. SQL implement data integrity rules whereas, NoSQL permits any data to be saved anywhere at any time without any verification. In terms of Query processing, they are almost the same.

Overall, choose the best among them, it all depends on the user, on the needs of the database and the demand of the query.

Hadoop and MapReduce

As mentioned above Hadoop is an open source ecosystem composed of multiple components to address various use cases in the big data ecosystem. The open-source software essentially means Hadoop is free and there is no need for the license, it is free of cost whereas, Map Reduce is a programming model for large scale parallel processing of Data. Map reduce is an execution model in a Hadoop framework and it processes large data in parallel. Hadoop is the platform for MapReduce program because Hadoop allows its code to process on data, which is mapping and reducing. Hadoop provides fault tolerance, which is the replication of block data and speculative execution for MapReduce job. Map Reduce is nothing but an algorithm whereas, Hadoop is an Ecosystem and they both are independent of underlying hardware or programming language.

Big Organizations and Companies choose Hadoop and MapReduce for large data such as Netflix's using Hadoop for query processing and Business Intelligence, Yahoo! Search Webmap is using Hadoop for its 10,000 core Linux cluster which produces the tons of data and that data is used in every Yahoo! Web search query and Facebook uses largest Hadoop cluster in the world with 21 PB (petabyte) of storage. Hadoop's one of the issues is, it cannot Integrate with existing systems might prove to be difficult and the other issue is the security issue, when it comes to safe enterprise deployment, especially if it concerns sensitive data.

There are various programming language tools available and each of them has distinct functions. The tools which are mentioned above are the most used and powerful tools available.

10 Conclusion

The demand for "Big Data" is evident and is growing at a rapid pace, creating a new market and opportunities along with it. In this paper, detailed information on upcoming profile opportunities namely Data Scientist, Data Analyst, Data Architect, Statistician and Data Administrator is presented. After conducting primary and secondary research, skill sets, tools and certifications required in each position. Data professionals are required in almost every sector in today's world with specific roles and responsibilities ranging from database management, data development, data analysis and interpretation. Various other additional responsibilities are also presented in this paper. In recent decades presence of Big Data everywhere is also evident with organizations benefiting from achieving objectives of faster and better decision making, new products and services and cost saving, however with these immense advantages, Big Data possess the threat to the Privacy and Security. Quintilian bytes of data are generating each day and it raises the issue of Security and privacy. With more diversifies skill sets and opportunities and ever-growing demand of "Big Data" market trend is blooming rapidly and with further enhancement in its network and security, Big Data is set to create the humongous amount of jobs in the year to come.

References

[1] Barc. *Research.* Available at `http://barc-research.com/big-data-analysis-shown-to-increase-revenues-and-reduce-costs/`, version 1.6.0.

[2] Jeff Bezanson, Stefan Karpinski, Viral B Shah, and Alan Edelman. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.

[3] Lynne Billard. The role of statistics and the statistician. *The American Statistician*, 52(4):319–324, 1998.

[4] Andrew D Birrell. An introduction to programming with c# threads. Technical report, Technical report, Microsoft Corporation, May 2003. Manuscript available at http://research. microsoft. com/˜ birrell/papers/ThreadsCSharp. pdf, 2003.

[5] Terence Craig and Mary E Ludloff. *Privacy and big data.* " O'Reilly Media, Inc.", 2011.

[6] Thomas H Davenport and DJ Patil. Data scientist. *Harvard business review*, 90:70–76, 2012.

[7] Christopher Gandrud. *Reproducible Research with R and R Studio.* CRC Press, 2013.

[8] Rudy Hirschheim, Heinz K Klein, and Kalle Lyytinen. *Information systems development and data modeling: conceptual and philosophical foundations.* Cambridge University Press, 1995.

[9] Vilas Jadhav. Data mining of big data: The survey and review.

[10] A. Katal, M. Wazid, and R.H. Goudar. Big data: Issues, challenges, tools and good practices. In *Contemporary Computing (IC3), 2013 Sixth International Conference on*, pages 404–409, Aug 2013.

[11] Ralph Kimball and Margy Ross. *The data warehouse toolkit: the complete guide to dimensional modeling.* John Wiley & Sons, 2011.

[12] Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. Big data. *The management revolution. Harvard Bus Rev*, 90(10):61–67, 2012.

[13] AB MySQL. Mysql, 2001.

[14] Daniel Nunan and MariaLaura Di Domenico. Market research and the ethics of big data. *International Journal of Market Research*, 55(4):2–13, 2013.

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[16] Sarbjit Rai. Sas® certification: An end user's review.

[17] Tom White. *Hadoop: The definitive guide.* " O'Reilly Media, Inc.", 2012.

[18] Shuyun Xu, Yu Li, and Shiyong Lu. Erdraw: An xml-based er-diagram drawing and translation tool. In *Computers and Their Applications*, pages 143–146, 2003.