Big Data Careers

Krishna Kumar

Abstract - Increasing number of organizations are discovering that running fast on a Big Data Journey is not an easy task. On one hand there is a dearth of talent on the latest tools and technologies but on the other hand their existing business experts lack requisite appreciation of the potential / capabilities of big data and its potential and know how around to tap into it in a systematic manner. This article discusses some of the career profiles in the Big Data domain.

Many economists in the advanced economies of the World believe that "Eventually Data will surpass Crude Oil, in importance."

In essence, the success of organization in leveraging the Big Data dream essentially boils down to its ability to put together a crack team formed from empowered people with diverse skills who can collaboratively work together to make meaningful business assumptions, take incremental risks, conduct meaningful experiments, learn to implement and scale up the successful implementations.

Objectives of this core team differ, depending on the strategic imperatives being chased by the organization. Based on these, the skills sets needed in the team also changes significantly from assignment to assignment. Because of this, the Big data Roles are continuously evolving.

Some of the high level objectives being pursued by the Big data teams are as below.

- Recognition (image, text, audio, video, gestures, facial expressions.)
- Scoring / Ranking – (FICO Score)
- Segmentation (Demographic based marketing)
- Forecasts (e.g. sales/revenue/ expenses)
- Optimization (Risk management)
- Prediction (predict value based on some inputs)
- Classification (this bucket or that)
- Recommendation (Shopping carts add ons)
- Anomaly detection (Frauds)
- Pattern detection / Grouping (classification without known causes – buying combos)

To achieve some of these common goals as above, there are different roles that are gaining prominence in this world of big data. They are listed below.

1 Chief Data Officer

This is akin to a CXO position, reporting directly to the board or to the CEO of the organization. The Chief Data officer is Responsible for overall big data strategy within an organization. – ensuring that the data is accurate, secure and customer privacies are governed correctly.

He/she is responsible for Crafting and managing standard data operating procedures, data accountability policies, data quality standards, data privacy and ethical policies….and be able to understand how to combine diff data source (*from within and outside*) with each other.

In short – the Chief Data Officer, manages the data of the organization as a strategic asset.

1.1 Key Skills

- Broad appreciation of Business Goals & how different data sources/ types enables them
- Strong Leadership and Communication Skills to interact with the board and senior business leaders, effectively
- Experience in leading major Information management programs in key business areas
- Expertise in creating and deploying best practices and methodologies
- Demonstrate Policy thinking
- Knowledge of developing business cases for technical projects with lot of uncertainities
- Familiarity with Modelling techniques, predictive modelling and big data tool sets

This is a new role that has evolved in the last couple of years and different organisations are actively evaluating this specialized role and its suitability for their organization. *20+ years of experience in not unu-*

sual for this position.

2 BIG DATA SOLUTION ARCHITECT

A skilled architect with cross industry, cross-functional and cross-domain know how. He sketches the big data solution architecture, and monitors and governs the implementation of the same.

He puts the discovered data in such a organised form so that it can be analysed. He/She structures the data so that they can be usefully queried in appropriate timeframes by different users. They ensure data updation happens in a predetermined manner for it to continuously remain useful.

2.1 Key Skills

- Experience in having designed normal solution architecture before coming into the big data solutions space (*15+ years of experience is very normal for this position)*
- Experience in architecting large Data warehouses with good understanding of Cluster / Parallel architecture as well as highscale distributed RDBMS / NoSQL Platforms is important
- Experience on Cloud computing infrastructure like Amazon Web Services, Elastic MapReduce, Azure etc.
  - Experience in major big data solutions like Hadoop / Mapreduce, Hive, Hbase, MongoDB, Cassandara
  - Depending on the project experience in Impala, Mahout, Flume, ZooKeeper/Sqoop are important
- Firm understanding of major programming/ Scripting language – Java, R, PHP, Linux, Ruby, Python.
- ETL Tool experience – Informatica, Talend, Pentaho
- Knowledge of data security, data privacy etc.
- Capabilities
  - Articulate Pros /Cons of various options
  - Benchmark systems, analyse system bottlenecks and propose solution to eliminate them
    - Document complex use cases, solutions and recommendations
    - Work in fast paced agile environments

3 DATA SCIENTIST

It is probably the most talked about Job Profile in the world today. Renowned as the "Sexiest Job of the 21st century " - as published by HBR.

This role is a very demanding role wherein the person playing this role is to have a deep appreciation of business domain combined with an ability to statistically appreciate the nature and variety of data and have a technical capability to leverage different technologies/ tools to deal with the deal with data in such a way as to guide the business to take decisions which might lead to solving the business problem at hand.

In short a Data scientist is a Business analyst, data modeler, a statistician and a developer all rolled into one.

Typically, a Data Scientist is familiar with the Business domain and the datasets accompanying it. He/ She creates sophisticated analytical models, that help solve a business problem – for e.g. pricing optimization across channels, predict customer behavior etc.

3.1 Key Skills

- Business Domain
  - Marketing, consumer behavior, Supply chain, finance, healthcare etc.
- Statistics / Probability
  - R
  - Correlation, baysian clustering,
  - Predictive analysis
- Computer Science / Software programming
  - Languages – Java, Python
- Written / Verbal Communication Skills
- Technical Proficiency
  - Database systems such as MySQL, Hive etc.
  - Data Mining, Machine learning (Mahout)

4 BIG DATA ENGINEER

Sources of data are ever expanding – different types of data –files, text messages, images, audio, video, gestures – from different kinds of sources such as application data, reports, social sites, sensors etc.

Based on the solution provided by the Big Data Solution Architect, a data engineer determines the way to tap into the various kinds of useful data from different sources, how to bring it into the organization (Builds data pipes into the organization), how to store them, retrieve them, combine them and serve them for the use of the different stakeholders like data scientists, machine learning scientists, analysts etc. and determines how to archive it, retire the data etc.

Big Data Engineers builds large scale data processing systems and algorithms. They typically analyse each source of data and determines the kind of pipe they need to setup for that specific data (depending on the complexity of the different Vs – Volume, value, veracity, volatility, velocity, variety etc), cleanse it, get it ready for processing and serve it to the different stakeholders. He also develops strategies for staging and archiving data.

Also, in a way – the data engineer is also a "Data hygienist" – who ensures that the data coming into the system is clean and accurate and ensures that the data stays that way throughout the data life cycle.

4.1 Key skills

- SQL & relational databases like Redis
- NoSQL Databases like MongoDB
- Apache Hadoop and its ecosystem – MapReduce, Hive
- Apache Spark and its ecosystem
- Languages – Scripting language like Java, C++, Ruby, Python & R

5 Machine Learning Scientist/ Engineer

The machine learning scientists are those who are involved in crafting and using predictive and correlative tools used for leverage data.

They work in the R&D of algorithms that are used in adaptive systems. They build methods of predicting product suggestions and demand forecasting and explore Big data automatically extract patterns etc.
In many situations, the machine learning engineer's final "Output" is the working software and their audience for this output consists of other software components that run automatically with minimal human supervision. The decisions are made by machines and they affect how a product or service behaves.

Machine learning Scientists creates algorithms allow for application of statistical analysis at high speeds. They design interrogation of data with enough statistical understanding to know that when the results are not to be trusted.

Statistics and Programming are the 2 biggest assets to the machine learning practitioner.

5.1 Key Skill areas

- Statistics
- Intermediate level Algebra/ calculus
- Programming skills – C++, Python
- Learning theory (intermediate level)
- Understanding of the inner workings of the arsenal of machine learning algorithms

6 Big Data Analysts

A big data Analyst primarily works with data in a given system and performs analysis of the given data set. He helps the data scientist in performing the necessary jobs. Many times Analysts graduate to don the role of data scientists after they gain valuable experience in the analyst role.

6.1 Key Skills

- Business acumen
- Should enjoy – discovering , solving problems
- Data Mining (Data auditing, aggregation, validation, reconciliation)
- Advanced data modelling
- Testing
    - A/B testing on different hypotheses – to directly/indirectly impact Key Performance indicators (KPIs)
- Creating clear/concise reports to explain results
- Technical Skills
    - SQL databases
    - BI platforms – tableau
    - Basic knowledge of Hadoop/ MapReduce
    - Statistical packages – R, Matlab, SPSS
    - Programming Languages

7 Campaign Specialists

They are business folks, who use the output of all the big data analytics, pilot it in the real world. They take these models and translate this into specific campaigns with the target audience to drive the business results. In a way, they help translate these models to business results.

During the piloting phase, they gain first- hand knowledge and document their learnings. If they outcomes are beneficial / valuable, they fine-tune their approach to be adopted and then scale it up for an organization wide implementation.

In short, These folks are often experts in the functional area and with the behavior of the target segment. They therefore define business steps which has

the targeted items based on the model and ensure that there are follow through action items to ensure that the implementation progresses in a campaign like manner.

7.1 Key Skills

- Strong Domain Skills in the relevant area – to operationalize the experimentation
- Determine measures of success / failure
- Understanding of data models
- Strong communication skills to train final users and articulate the correct way to leverage the findings and make them apply the right method for their specific data sets.

8 Big Data Visualiser

In Big data, one of the key ability is to visualize the data, such that the senior management in the organization can appreciate it, play with it to find new patterns and insights.

Instead of usual graphs/ pie charts etc. the data visualizer can help tell a compelling story or insights through a mixture of interactive visuals to deliver the insights. A data visualizer has the necessary skills to turn abstract information from data analytics into appealing and understandable visualisations that clearly explains the results of the analyses.

8.1 Key skils

- Creative thinker - who understands UI/UX and has visualizing skills such as Typography, interface design, visual art design
- Programming skills to build visualisations
- Good background in Source control, testing frameworks as well as agile development practice
- Use metadata, metrics, colors, size, position to highlight
- Technical Skills
  - JavaScripts, HTML, CSS, R
  - Modern Visualisation Frameworks such as Gephi, Processing, d3js,
  - Web libraries such as JQuery, LESS, Functional Javascript
  - Photoshop, Illustrator, Indesign
  - Excellent Written / Communication

9 Big Data Programmer

The canvas for learning programming can be across various stages through which big data is sourced, processed and used.

The programmer roles are available in all the stages of the Big data journey from its source, to the models to the visualization of results and rolling out the solution on an organization wide scale.

Big data programming involves various tools / languages that are used by the various role players like Big data architect, the data engineer, ETL specialists, Modellers, Analysts, Visualizers etc.

9.1 Key Skills

- Programming languages like C++, Python, R
- Reporting tools – Jasper reports, Kibana, Tableu, SAS etc
- Big data tools – Hbase, MapReduce, Python, Hive, Spark etc.
- Frameworks such as Elastic Search,
- Ability to explore, self- learn, share knowledge, collaborate effectively in teams
- Work with large data sets and understand the nuances of different types of data and sources will be of great help

These roles are the ones usually available to fresh graduates or for the first time entrants into the big data spares. To gain proficiency on any of these tools, there are various online courses available which can kick-start their journey into this space. And depending on their area of interest and acumen, the youngsters can take seek to become statisticians, data analysts, data engineers etc. by learning others skills on the job. They can also choose to become specialists on different kinds of data and carve a specialist career for themselves.

Krishna Kumar Thiagarajan received his B.Tech from NIT Surat in 1991, MBA from SP Jain, Mumbai in 1996, CFA from ICFAI, and a Business Leadership diploma from U21 Global. He is a frequent speaker on contemporary IT topics such as Big Data, IoT, and Smart City with Hyderabad University and Digital India. He has varied interests and is into designing dashboards and is a co-producer of a Marathi film, "Satrangi Re". He is certified Six Sigma black belt holder. Krishna Kumar received International HR Leader award from USA and has certificates in Big Data from University of California and in Executive Data Science from John Hopkins University, USA.