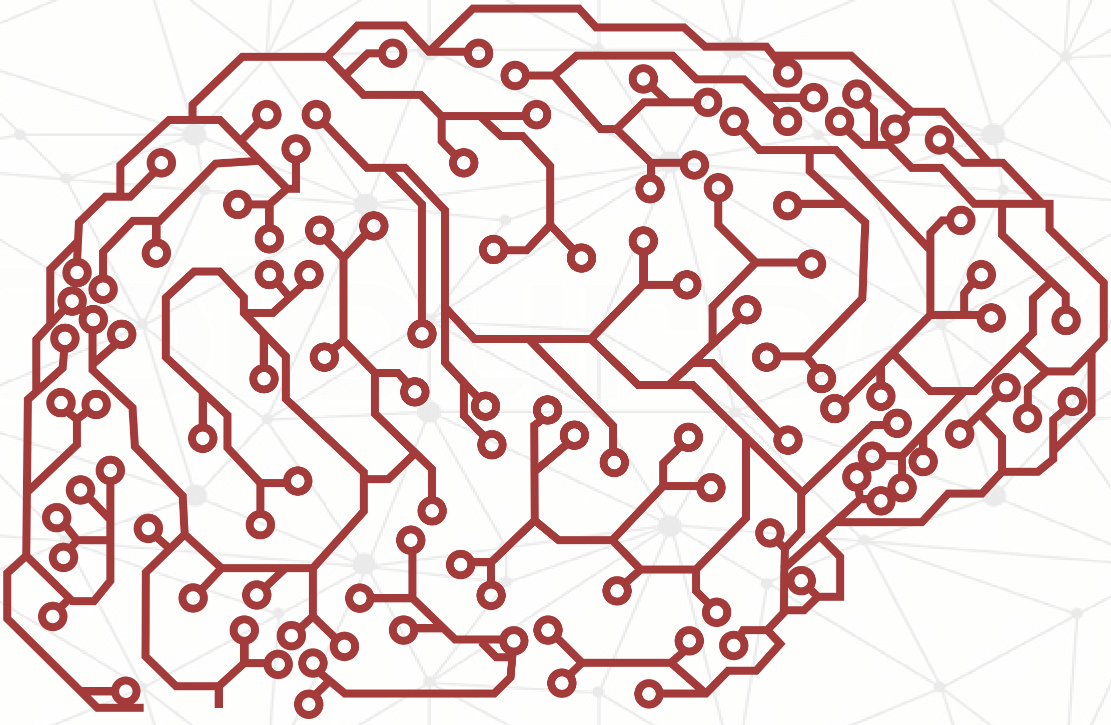
ACADGILD

STEP BY STEP GUIDE TO BECOME BIG DATA DEVELOPER



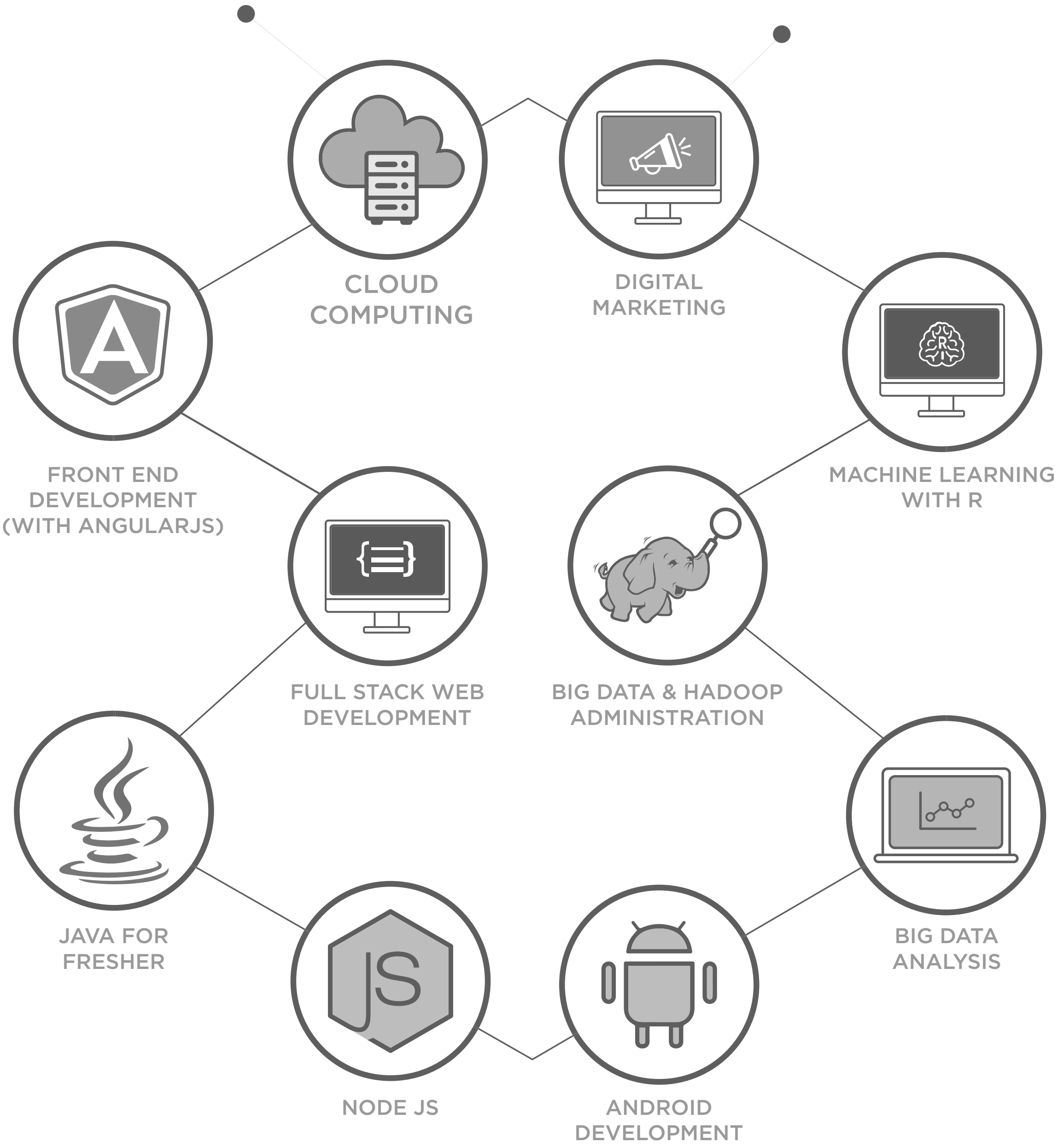
LEARN.DO.EARN

acadgild.com

About ACADGILD

ACADGILD is a technology education startup that aims to create an ecosystem for skill development in which people can learn from mentors and from each other. We believe that software development requires highly specialized skills that are best learned with guidance from experienced practitioners. Online videos or classroom formats are poor substitutes for building real projects with help from a dedicated mentor. Our mission is to teach hands-on, job-ready software programming skills, globally, in small batches of 8 to 10 students, using industry experts.

ACADGILD offers courses in



Watch this short video to know more about ACADGILD.

Disclaimer

This material is intended only for the learners and is not intended for any commercial purpose. If you are not the intended recipient, then you should not distribute or copy this material. Please notify the sender immediately or click here to contact us.

Published by
ACADGILD,
support@acadgild.com

TABLE OF CONTENTS

STEP-BY-STEP GUIDE TO BECOME A BIG DATA DEVELOPER

Introduction

In current job scenario, where Big Data is one of the most sought after skill in IT industry, learning Big Data skills can surely open door to a highly lucrative career.

But the big question that boggles everyone's mind is, "How to become a Big Data Hadoop Developer?"

In this eBook, we will be discussing the best practices and steps, which will help you become **a Big Data/ Hadoop Developer.**

No one is denying the fact that Hadoop and Big Data have become synonymous to each other as Big Data processing has become easier than ever, with technologies like Hadoop and Spark.

Hadoop related jobs is no longer confined to just Tech companies but also other types of companies which includes, financial firms, retail organizations, banks, healthcare organizations, Government organizations, Advertisement sectors, etc. and recruiter are looking for Hadoops to work in these sectors.

So let's take a look at the step-by-step approach to become a Big Data Developer.

Big Data is a cluster of many technologies and tools that are used in various setups an fulfilling any of the below three skills sets will give the learner an upper hand to begin their career in Big Data.

Skills to become a Big Data Developer

Let's have a look at the necessary skills to become a Big Data Developer.

Step1: Prerequisites for Learning Big Data

To process to the next step of learning Big Data, you need to have at least one of the below mentioned skills, as these skills will you learn Big Data concepts easily.

Programming Skills:

Having the necessary programming skills is very vital and the first step to start with Big Data. Knowledge of following language will be great way to kick start your Big Data learning.

Java

If you are looking for a "Big Data Developer Job," learning Java is highly suggested. Big Data technology like Hadoop is written in Java; hence, the knowledge of Java basics is helpful to learn Hadoop.

Python

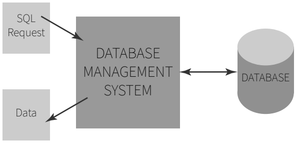
Python is considered to be the easiest programming language in the world because of its simple syntax, allowing you to learn it quickly.
Refer our blog series on Python tutorials to get started with Python

Unix/Linux Operating System and Shell Scripting

Good practice in shell scripting makes your life easier when it comes to Big Data. Many tools in Big Data has the command line interface where the commands are based on the shell scripting and Unix commands.

SQL (Structured Query Language)

SQL, popularly known as 'sequel', makes Hive (a query language for Big Data) easier. Playing around with SQL in relational databases helps us understand the querying process of large data sets.



Step 2: Learning about Big Data Platforms

Once you are confident that you will be able to solve basic problems using Java/Python/SQL, you are ready to take the next step. You need to learn about some Big Data technologies like Hadoop and Spark.

Hadoop would be the best place to start, as it is quickly becoming the nucleus of Big Data solutions for many enterprises. The Hadoop ecosystem is growing with lots of new technologies/stacks added every day. It would be really helpful for a professional to get familiar with the below technologies before starting their career in Big Data.

MapReduce

Is the Google paper that started it all (**Page on googleusercontent.com**). It is a paradigm for writing distributed code inspired by some elements of functional programming. You don't have to do things this way, but it neatly solves a lot of problems we try to solve in a distributed way. The Google internal implementation is called MapReduce and **Hadoop** is its open-source implementation. Amazon's Hadoop instance is called Elastic MapReduce (**EMR**) and has plugins for multiple languages.

HDFS

Is an implementation inspired by the **Google File System** (GFS) to store files across a bunch of machines when it is too big for one. Hadoop consumes data in HDFS (Hadoop Distributed File System).
To get an overview of HDFS Please go through the below blog on Introduction to HDFS

Hive and Pig

They are abstractions on top of Hadoop designed to help analysis of tabular data stored in a distributed file system (think of excel sheets too big to store on one machine). They operate on top of a data warehouse, so the high-level idea is to dump data once and analyze it by reading and processing it instead of updating cells, rows and columns individually. Hive has a language similar to SQL while Pig is inspired by Google's **Sawzall** (**Google Research Publication: Sawzall.**) You generally don't update a single cell in a table when processing it with Hive or Pig.

Refer the below blog on Beginner's guide for Hive to get started with Hive

Refer the below blog on Beginner's guide for Pig to kick start your learning in Pig

Mahout

(**Scalable machine learning and data mining**) is **a collection of machine learning libraries written in the MapReduce paradigm,** specifically for Hadoop. Google has its own internal version but they haven't published a paper on it as far as I know.

Oozie is a workflow scheduler

The oversimplified description would be that it's something that puts together a pipeline of the tools described above. For example, you can write an Oozie script that will scrape your production HBase data to a Hive warehouse nightly, then a Mahout script will train with this data. At the same time, you might use Pig to pull in the test set into another file and when Mahout is done creating a model you can pass the testing data through the model and get results. You specify the dependency graph of these tasks through Oozie (I may be messing up terminology since I've never used Oozie but have used the Facebook equivalent).

□ Solr

uses Lucene under the hood to provide a convenient REST API for indexing and searching data.

□ ElasticSearch

Is similar to Solr.

Sqoop

Enterprises that use Hadoop often find it necessary to transfer some of their data from traditional relational database management systems (RDBMSs) to the Hadoop ecosystem. Sqoop, an integral part of Hadoop and can perform this transfer in an automated fashion. Sqoop uses a connector-based architecture that allows it to use plugins to connect with external databases.

Refer this blog to get the deep understanding regarding the functioning of Sqoop

The NoSQL database

Also called as ‘Not Only SQL’, NoSQL is an approach to data management and database design that's useful for very large sets of distributed data. This database system is non-relational, distributed, open-source and horizontally scalable. NoSQL seeks to solve the scalability and Big Data performance issues that relational databases weren’t designed to address.

AMost popular NoSQL databases in Big Data projects these days are **Apache Cassandra, MongoDB and HBase.**
To get more insights on Big Data terminologies, you can refer to this post.

Learning Spark

Now comes the next part of your learning process – learning Spark. This should be done after gaining a little bit of experience with Hadoop. Spark will provide you the speed and tools that Hadoop couldn't. However, in order to learn Spark, you need to have prior knowledge of Scala/Python/Java/R to use it.

We suggest our readers to go through the below links which would help them to start learning about Spark.

https://acadgild.com/blog/beginners-guide-for-spark/

https://acadgild.com/blog/rdd-spark/

https://acadgild.com/blog/introduction-spark-rdd-basic-operations-rdd/

Step 3: Learning Analytics & Visualization:

After solving the Big Data problem, we acquire data in a manageable format, and platform is setup for generating reports. Most enterprise architectures leveraging Hadoop still have a SQL Database to store and report data out of Hadoop. Loading data out of Hadoop and into a SQL database is a good practice in the real world, but for the sake of learning the Big Data side of it, it is not necessary. Several (free) reporting tools are available out there, that will connect to Hadoop/Hive directly and will work fine for learning purposes.

Many visualization tools like Tableau, QlikView and Power BI are very much popular in the industry.

TOP RESOURCES FOR BIG DATA HADOOP DEVELOPER

Following are the top resources which can be valuable to Big Data Hadoop developers.

BOOKS

Hadoop The Definitive Guide

– by Tom White

This is the best book for Hadoop beginners. This book can you adapt to the world of Big Data management. The book is written in an easy language which can be understood by a beginner. The first eight chapters are critical in understanding the nuances of this technology. A basic knowledge of Java language is an advantage and helpful in understanding the book better.

Knowledge of Hadoop administration is useful in understanding how the framework works internally.

Hadoop Operations

– by Eric Sammer

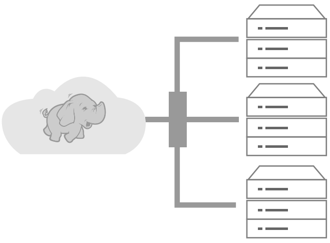
This book discusses the technology in depth, detailing the particulars of running Hadoop in actual production. The chapter on HDFS is very useful and this book can well qualify as a user's manual where all aspects of Hadoop technology are explained in an easily understandable language. The book would impress the beginners and advanced learner as well because all the aspects of software and related technologies are explained in great detail.

Lucene

Tis a bunch of search-related and NLP tools but its core feature is being a search index and retrieval system. It takes data from a store like HBase and indexes it for fast retrieval from a search query.

Professional Hadoop Solutions

by Boris Lublinsky, Kevin T. Smith, Alexey Yakubovich



This book is much more than just an overview or a guide and provides in-depth insight about Hadoop and related technologies. The first chapter is a sort of an introduction and as you proceed further on the various aspects of MapReduce programming, Oozie, you would realize how critical it is to understand these technologies to become a skilled Hadoop developer. This book is recommended for all advanced learners of Big Data management technology and those who want to apply the concepts in their own domain.

If you want to get an in-depth knowledge about specific tools, then it is recommended to read the books like:

- □ Programming Pig by Alan Gates
- □ Programming Hive by Dean Wampler, Edward Capriolo, and Jason Rutherglen
- □ MapReduce Design Patterns by Adam Shook and Donald Miner

BLOG

The following blogs are useful and provide information on various aspects of Hadoop along with the latest updates.

Hadoop Weekly

Hadoop Weekly is a summary of the week's top news in the Apache Hadoop™ ecosystem. It is aimed at developers or those trying to keep themselves updated with the latest Hadoop developments.

It gathers the latest and the most shared Hadoop content in Twitter, Facebook, Google+ and LinkedIn every week. By subscribing to this weekly channel, you won't miss any important new Hadoop content. Guaranteed! Join 7,428 other subscribers today.

https://www.hadoopweekly.com/index.html

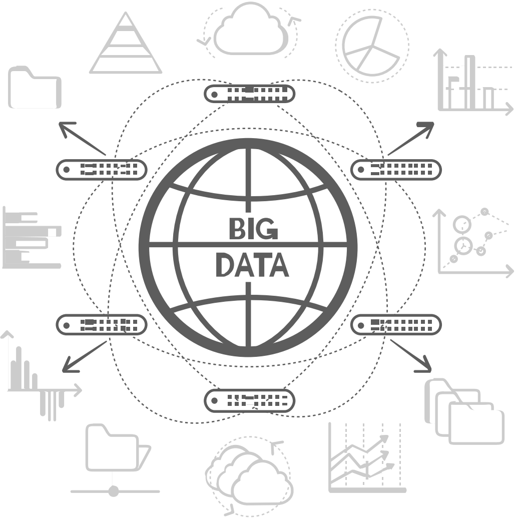
AcadGild Blog

AcadGild is a technology education start-up which provides online courses in latest technologies such as Android development, Big Data, Cloud computing, Front End development, Digital Marketing, Machine Learning, Node JS, Robotics.

https://acadgild.com/blog/

Data Science Central

Data Science Central is the industry's online resource for big data practitioners. Ranging from Analytics to Data Integration and Visualization, Data Science Central provides a community experience that includes a robust editorial platform, social interaction, forum-based technical support, the latest in technology, tools and trends, including job opportunities.



The Hortonworks Blog

Hortonworks is a leading innovator in the industry, creating, distributing and supporting enterprise-ready open data platforms and modern data applications. They have a very good repository of Hadoop and related technologies.

Cloudera Engineering Blog

Cloudera provides a modern platform for data management and analytics. It provides the world's fastest, easiest, and most secure Apache Hadoop platform. This helps you solve the most challenging business problems with data.

You can download Hortonworks or Cloudera Sandbox or Apache Hadoop AcadGild VM.

Cloudera QuickStart VM

Hortonworks Sandbox

Data Analytics

You can get information on data mining from the following link.

Data Mining – Research at Google

You can visit the official Apache Hadoop website for more information and latest news on Hadoop.

http://hadoop.apache.org

LINKEDIN GROUPS

Hadoop Users

This group is the original and most established group of Hadoop users on LinkedIn.

Description: A group for Hadoop users

Activity: Active, 47 discussions this month

Member Count: 24,236 members

Created: October 7, 2008

Cloudera Hadoop Users

This group is a subgroup of the above Hadoop Users group, focusing on Cloudera, which is a commercial enterprise version of Hadoop. Below is a description about the group.

Description:

Cloudera (www.cloudera.com) is the leading provider of Apache Hadoop-based software and services and works with customers in financial services, web, retail, telecommunications, government and other industries. Cloudera's Distribution for Apache Hadoop and Cloudera Enterprise, help organizations with their rich source of information.

Activity: Very Active,117 discussions this month

Member Count: 4,033 members

Created: June 26, 2009

Hadoop Hive

This group discusses the use of the Hive language, which is a high level SQL-like language used on top of MapReduce for easy access to data on Hadoop.

Description:

Hive is a data warehouse infrastructure built on top of Hadoop that provides tools to enable easy data summarization, ad-hoc querying and analysis of large dataset stored in Hadoop files. It provides a mechanism to enforce structure on this data and it also provides a simple query language called Hive QL which is based on SQL, which enables users who are familiar with SQL to query this data. At the same time, this language also allows traditional map/reduce programmers to be able to plug in their custom mappers and reducers to do more sophisticated analysis which may not be supported by the built-in capabilities of the language.

Activity: Very Active,70 discussions this month

Member Count: 3,339 members

Created: June 15, 2009

Notes: It is a subproject of Hadoop

Hadoop India

This group is on the top 10 list and consists of regional members in India who use Hadoop. The most represented area in this group is Bengaluru, India (29%).

Description:

Group for Hadoop India Users

Activity: Very Active,199 discussions this month

Member Count: 3,185 members

Created: September 2, 2009

Visit the group on LinkedIn

BIG DATA PROFESSIONALS, ARCHITECTS, SCIENTISTS, ANALYTICS EXPERTS DEVELOPERS CLOUD COMPUTING NOSQL BI

Description:

This is an extremely active group, with over 900 new discussions per month. Remember, you can set your group email preferences to a summary digest to limit the number of emails that you receive from this group.

Technologies covered

Information Technology, Predictive Modeling, Business Intelligence (BI), Decision Support, Text Mining, Machine Virtualization, Statistics, Apps Developer Software Enterprise, Mobile Web Oracle Database, SAAS, Linux, Java API, Cloudera, MapR, Greenplum.

Activity: Very Active, 937 discussions this month

Member Count: 14,828 members

Created: Created: September 1, 2008

Visit the group on LinkedIn

Check out these resources to enrich your skills in Big Data and Hadoop.

We hope this eBook has been helpful in understanding the vital steps necessary to make your career in Big Data domain.

This EBook was developed with inputs from **Abinav Sharma,** Product Designer at Quora.

For a better understanding and in-depth learning of Big Data and Hadoop technology, enroll for our **Big Data and Hadoop Development** course.

Keep visiting our websites **www.acadgild.com** for more posts on Big Data and other technologies.

ACADGILD

Check out these resources to enrich your skills in Big Data and Hadoop.

We hope this eBook has been helpful in understanding the vital steps necessary to make your career in Big Data domain.

For a better understanding and in-depth learning of Big Data and Hadoop technology, enroll for our Big Data and Hadoop Development course.

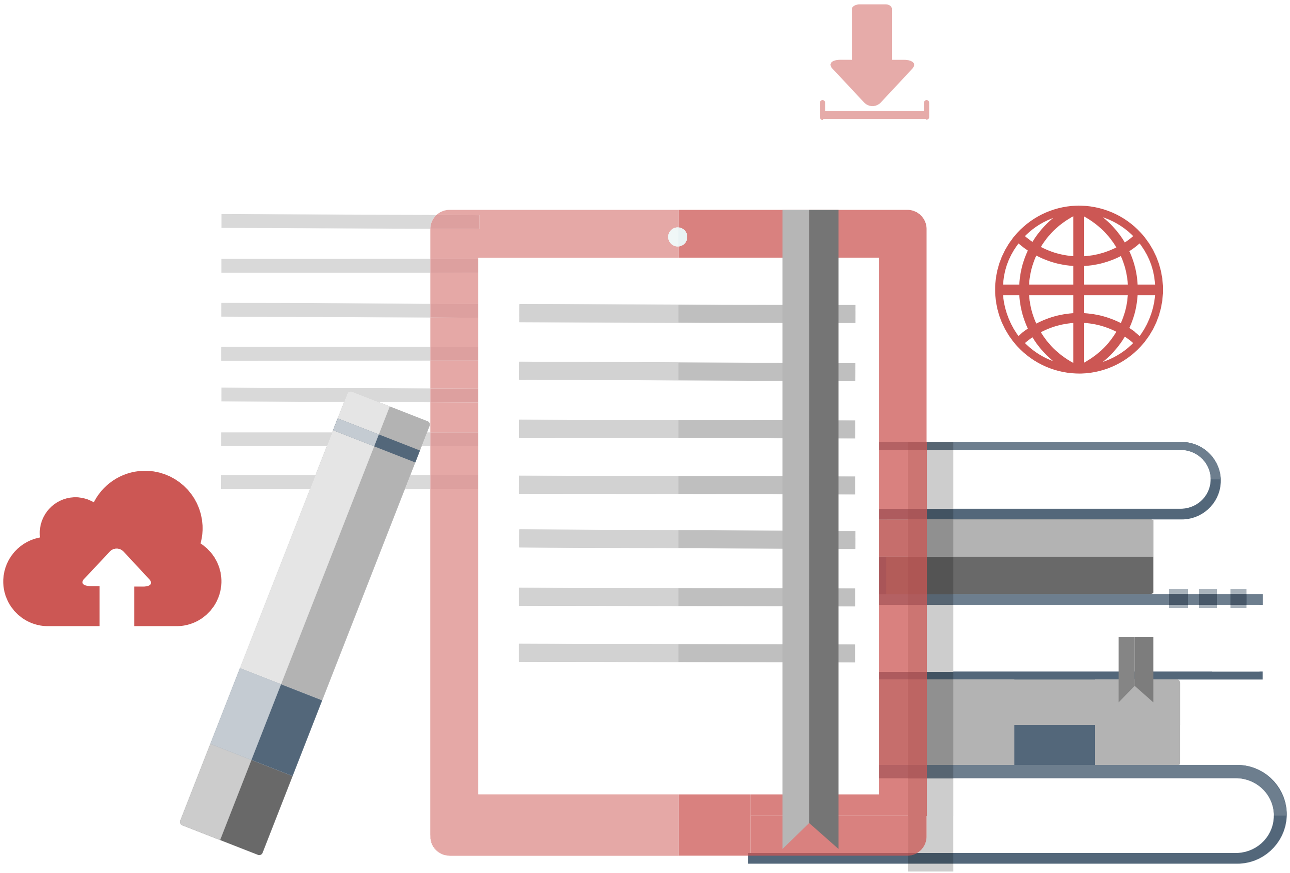
Keep visiting our websites www.acadgild.com for more posts on Big Data and other technologies.

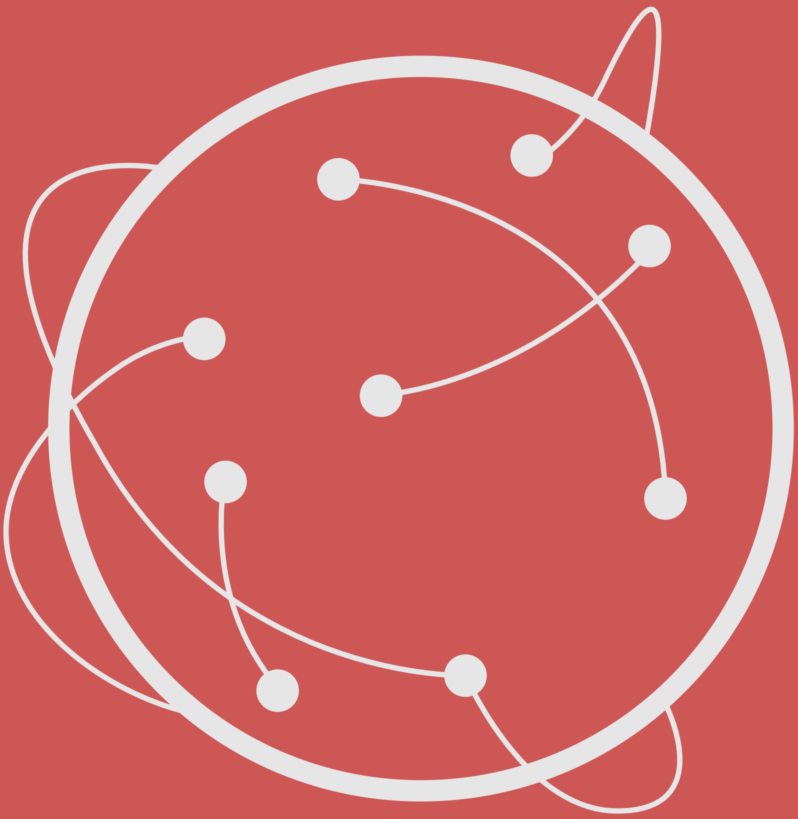
ABOUT THE AUTHOR

Satyam Kumar is a Big Data Professional, working in AcadGild with 3+ years of experience and having expertise in Big Data technologies like Hadoop, Spark, NoSQL and other related technologies.

He strives to code in Programming languages like Java and Python and have been responsible for development of various projects and blogs related to Hadoop ecosystem and Spark.

Feel free to contact him at **satyam@acadgild.com** in case you have any query.





ACADGILD

We hope this Ebook has helped you understand some of the terminology associated with Big Data. If you have any questions, feel free to contact us at support@acadgild.com.