International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Analyze Human Genome Using Big Data

Poonm Kumari[1], Shiv Kumar[2]

[1]Mewar University, Chittorgargh, Department of Computer Science of Engineering, NH-79, Gangrar-312901, India

[2]Co-Guide, Mewar University, Chittorgargh, Department of Computer Science of Engineering, NH-79, Gangrar-312901, India

Abstract: ***As day by day population increase or internet user's increases, our data also increase. One day we will be unable to handle these data manually or using data processing application or using handmade database management. Our data increases with the rate of 2.5 quintillion bytes per day. One quintillions means 10 [30]. So, we created 2.5 * 10 [30] * 8 bit per day. In similar fashion scientist data is increases day by day in many areas like genomics, biological research, meteorology, simulation in physics etc. So we need some tool to handle it properly like Hadoop, Crossbow etc. Because we are unable to search or store or transfer these large amounts of data. Hadoop uses map reduce method to handle it while Crossbow uses pipelined map reduced method to handle this data. So, performance increases and reliability also increases because result is obtained by analysis Big data which is much more important for biological research or genomics without knowing the implementation of parallelism at lower level.***

Keywords: Hadoop, Crossbow, Data Warehousing, Big Data, Human Genome, Map Reduce, Pipelined.

1. Introduction

All the animals differ because of DNA Structure. DNA contains all the information about the specie and history of its family. So to develop methods for sorting, retrieving, organizing and analyzing biological data, Bioinformatics is used. For new drug discovery and discovery of new disease which flows from one generation to next generation, Bioinformatics which is the melding of Molecular Biology and Computer Science is used. Complete set of DNA, including all of its genes containing all of the information needed to build and maintain that organism is called a Genome. In Humans a copy of the entire genome more than 3 billion DNA base pairs is contained in all cells that have a nucleus. So to store this much information and extracting useful patterns various techniques are used.

Cloud Computing provides solutions to deal with extremely large data and many applications are being developed to support the field of Bioinformatics using Hadoop and MapReduce. For example, the Cloudburst software is based on parallel read-mapping algorithm optimized for mapping next generation sequencing data to the human genome and other reference genomes.

Hadoop is an open-source flexible infrastructure for large scale computing and data processing on a network of clusters of commodity hardware. It supports big data distributed applications and allows applications to work with thousands of nodes maintaining reliability and fault tolerance ensures a robust system.

Hadoop is an Apache Software Foundation project being built and used by a global community of contributors, written in the Java programming and with the help of some scripts. It was originally derived from Google's MapReduce and Google File System by Doug Cutting [1].

Hadoop includes sub-projects such as:
Hadoop MapReduce: is used for processing and extracting knowledge from large data sets on compute clusters.

HDFS: is a scalable and distributed file system used for file storage. It supports a configurable degree of replication for reliable storage and provides high throughput access to application data. HDFS is inspired by the Google File System (GFS).

HBase: is a distributed database that supports storage of large tables and runs on top of HDFS. Pig and Hive are used for data analysis. Pig is a high level language running on top of MapReduce. It is an execution framework for parallel computing. Hive is running on top of Hadoop and provides database functionality.[2]

1.1 Hadoop Based on Map & Reduce

The easiest way to access external files or external data on a file system from within an Oracle database is through an external table. The data stored in external tables in a file system can be used transparently in SQL queries. These external tables could be used to access data stored in Hadoop File System from inside the Oracle database. Unfortunately HDFS files are not directly accessible through the normal operating system calls that the external table driver relies on. The FUSE (File system in User space) project provides a solution in this case. FUSE drivers that allow users to mount a HDFS store and treat it like a normal file system. By using one of these drivers and mounting HDFS on the database instance (on every instance if this was a RAC database), HDFS files can be easily accessed using the External Table infrastructure [3].

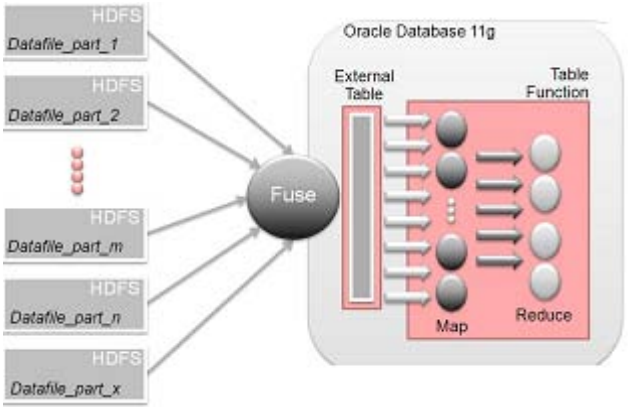

Figure 1: Accessing via External Tables with in-database Map Reduce [3]

In Figure 1 we are utilizing Oracle Database 11g to implement database mapreduce as described in this article. In general, the parallel execution framework in Oracle Database 11g is sufficient to run most of the desired operations in parallel directly from the external table.

If FUSE driver is unavailable then the external table approach may not be suitable. We have another option to fetch data from Hadoop using Oracle Table Functions.. At a high level we implement a table function that uses the DBMS_SCHEDULER framework to asynchronously launch an external shell script that submits a Hadoop Map-Reduce job. The table function and the mapper communicate using Oracle's Advanced Queuing feature. The Hadoop mapper en-queue's data into a common queue while the table function de-queues data from it. Since this table function can be run in parallel additional logic is used to ensure that only one of the slaves submits the External Job.

1.2 Crossbow

Crossbow is an open source software tool that uses two powerful algorithms Bowtie and SOAPsnp to give accurate results in short time with a little cost of computing. It works on Hadoop-enabled parallel processing to analyze human genomes in the clouds. It handles big data with the help of MapReduce method. If we take an example of sequencing of next generation genome Crossbow aligns reads and makes highly accurate SNP calls from a dataset comprising 38-fold coverage of the human genome in less than 1 day on a local 40 core cluster, and less than 3 hours using a 320-core cluster rented from Amazon's Elastic Compute Cloud3 (EC2) service. Crossbow's ability to run on EC2 means that users need not own or operate an expensive computer cluster in order to run Crossbow [4].

Bowtie: is a accurate genotyping algorithm with capacity of ultrafast, memory efficient short read aligner. It connects short DNA sequences called reads to the human genome at a rate of over 25 million 35bp-reads per hour. It indexes the genome with a Burrows-wheeler index to keep memory footprint small.

SoapSNP: within Hadoop to distribute and accelerate the computation. SNPs identified on the consensus sequence through the comparison with the reference genome.

Crossbow builds upon a parallel software framework called Hadoop. Hadoop is an open source implementation of the Map Reduce programming model that was first described by scientists at Google. Hadoop has become a popular tool for Computation over very large datasets, used at companies including Google, Yahoo, IBM, and Amazon. Hadoop requires that programs be expressed as a series of Map and Reduce steps operating on tuples of data. Though not all programs are easily expressed this way, Hadoop programs gain many benefits. In general, Hadoop programs need not deal with particulars of how work and data are distributed across a cluster or how to recover from failures. The insight behind Crossbow is that alignment and SNP calling can be framed as a series of Map, Sort and Reduce steps. The Map step is short read alignment, the Sort step bins and sorts alignments according to the genomic position aligned to, and the Reduce step calls SNPs for a given partition [4].

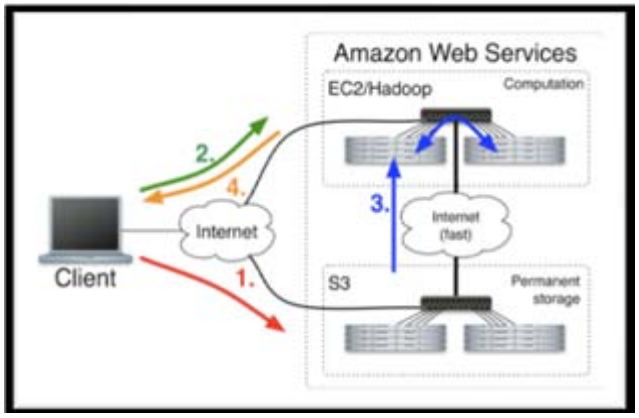

Figure 2: Steps involved in running Crossbow using Amazon's EC2 and S3 services [4]

1.3 Three-Tier Data Warehouse Architecture

In general data warehouses adopt 3-tier architecture. Following are the three tiers of data warehouse architecture [5].

Bottom Tier - The bottom tier of the architecture contains the data of warehouse database server. This is about the relational database in which we use back end tools to inset data into bottom layer. These back end tools and utilities perform all the operations like the Extract, Clean, Load, and refresh functions.

Middle Tier – OLAP Server is used at middle tier which can be implemented in following way:

By relational OLAP (ROLAP), which is an extended form of relational database management system?
The ROLAP maps the operations on multidimensional data to standard relational operations. By Multidimensional OLAP (MOLAP) model, this directly implements multidimensional data and operations.

Top-Tier – This is the front end layer which can be also called as client layer. Here we apply query tools and reporting tools, analysis tools and data mining tools.

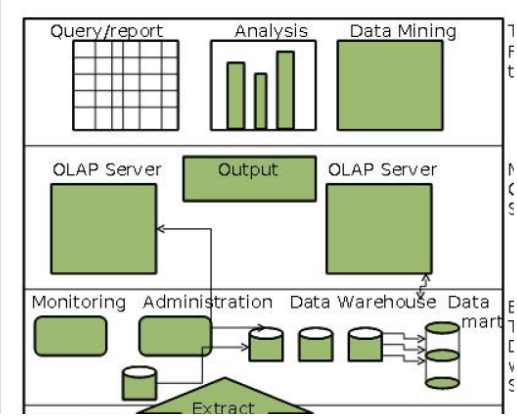

Figure 3: Three-tier Architecture of Data warehouse [5]

1.4 Human Genome

All type of genes of specie together known as Genome. Likewise Human Genome [6] is a complete set of genetic information about human. Genetic information includes information about a person genetic history, genetic tests, medical reports of past, present and future. The human genome comprises a sequence of approximately 3 billion component parts, called nucleotides, which are organized into DNA molecules—the double helix. The nucleotides, which serve as the alphabet for the language of life, are represented by just four letters: A, C, G, and T, corresponding to adenine, cytosine, guanine, and thymine. The nucleotide alphabet codes for the sequence of amino acids the body will use to build proteins

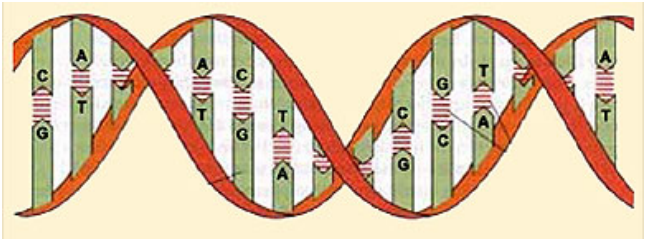

Figure 4: Human Genome [6]

Combinations of three nucleotides indicate one of twenty possible amino acids (for example, CCT codes for the amino acid glycine), so sets of nucleotide triplets form the instructions that cells use to build proteins. These proteins perform the work of the cells from development throughout life, contributing to both our physical attributes and many of our less tangible features, such as behavior, learning, and predisposition to disease. A segment of a DNA molecule that codes for one complete protein is called a gene. The human genome is carried on 23 different chromosomes—or DNA molecules.

Genomes of other species contain more or fewer nucleotides and chromosomes but follow the same basic organizational scheme as the human genome.

2. Handling Big Data Using Different -2 Tools

We can handle data or big data by using various tool like hadoop, RDBMS, Crossbow etc.

2.1 Handling Big Data Using Crossbow

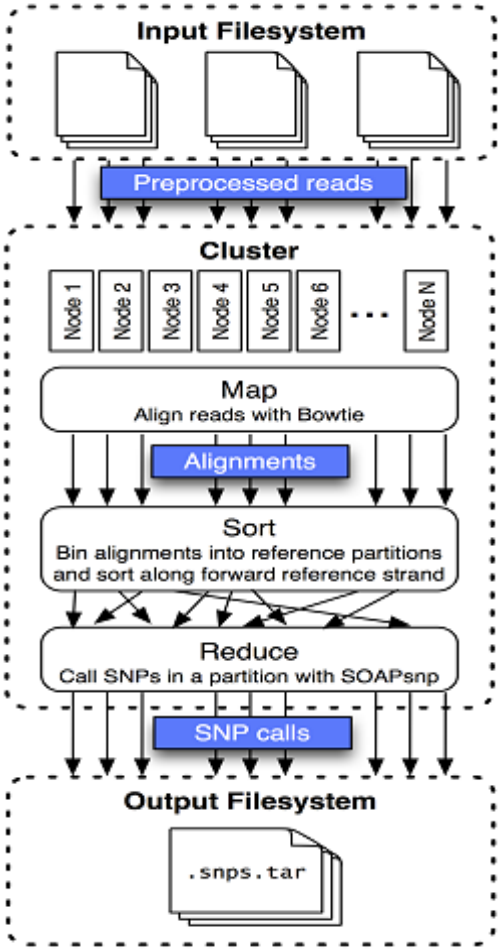

Figure 5: Crossbow Flow [7]

Description of above diagram [7]

The user first uploads reads to a file system visible to the Hadoop cluster. If the Hadoop cluster is in EC2 (Amazon's Elastic Compute Cloud), the file system might be an S3 bucket. If the Hadoop cluster is local, the file system might be an SNFS share.

A cluster may consist of any number of nodes. Hadoop handles the details of routing data, distributing and invoking programs, providing fault tolerance, etc.

- **Map** step is doing alignment of short reads. Many instances of Bowtie run in parallel across the cluster. Input tuples are reads and output tuples are alignments.
- **Sort** step binds alignments according to primary key (genome partition) and sorts according to a secondary key

(offset into partition). This is handled efficiently by Hadoop**.**

- **Reduce** step calls SNPs for each reference partition. Many instances of SOAPsnp run in parallel across the cluster. Input tuples are sorted alignments for a partition and output tuples are SNP calls.
- **Results** are stored in cluster's file system, then automatically archived and downloaded to the client machine. SNP calls are provided in SOAPsnp's format**.**

2.2 Handling Big Data Using Map Reduce [8]

Map Reduce is a programming model and software framework first developed by Google (Google's Map Reduce paper submitted in 2004). Intended to facilitate and simplify the processing of vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner. Map Reduce supports Petabytes of data or thousands of nodes on multiple clusters. The computation occurs on both unstructured and structured data.

Hadoop Simple Cluster Graphic

Figure 6: Hadoop simple cluster graphic [8]

Map Reduce core functionality
It provides facility to write user own analysis code either in java or other languages with the Hadoop Streaming API. MapReduce has two fundamental steps:

1) Map step:

Very first Master node takes problem input and breaks the input into sub problems then distributes these to worker or slave nodes. Worker nodes if sub problem input is enough large then breaks it again and distribute it to the lower level worker nodes. Likewise it forms a multi-level tree structure. These worker nodes process on small sub problems and hands back the result to Master node.

2) Reduce step:
In Reduce step master node collects the solutions from worker nodes and combines them in predefined manner to get the output to the original problem.

a) Data flow beyond the two key steps (map and reduce):
- Input reader – divides input into appropriate size splits which get assigned to a Map function
- Map function – maps file data to smaller, intermediate<key, value> pairs.
- ▫Partition function – finds the correct reducer: given the key and number of reducers, returns the desired Reduce node.
- Compare function – input for Reduce is pulled from the Map intermediate output and sorted according to this compare function
- Reduce function – takes intermediate values and reduces to a smaller solution handed back to the framework,
- Output writer – writes file output

b) A Map Reduce Job controls the execution
- Splits the input dataset into independent chunks
- Processed by the map tasks in parallel

3) The framework sorts the outputs of the maps
4) A Map Reduce Task is sent the output of the framework to reduce and combine
5) Both the input and output of the job are stored in a file system
6) Framework handles scheduling
7) Monitors and re-executes failed tasks

2.3 Comparing Hadoop and RDBMS

Figure 7: Comparing of Hadoop and RDBMS [9]

With the above solution it is clear that performance suffers because of traditional architecture where we have SAN or a NAS that is connected to a database with a bunch of applications connected to it and data is being constantly moved to where processing needs to happen. Such a model does not provide high performance at scale. What you really need is a distributed storage as well as processing platform such as Hadoop, where the functionality is run locally on the data, and the system scales linearly to extreme limits - even to geographically dispersed locations.[9]

2.4 Comparing Hadoop and Data Ware Housing [10]

Hadoop and the data ware house will often work together in a single supply chain. When it comes to Big Data, Hadoop excels in handling raw, unstructured and complex data with vast programming flexibility. Data warehouses also manages big structured data ,integrating subject areas providing interactive performance through B1 tools .It is rapidly becoming a symbiotic relationship .Some differences are clear ,and identifying workloads or data that runs best on one or the other will be dependent on any organization and use cases. As with all plate from selection, careful analysis of the business and technical requirement should be done before plate from to ensure the best outcome .Having both Hadoop and a data warehouse onsite greatly helps everyone learn when to use which.

Requirement	Data Warehouse	Hadoop
Low latency, interactive reports, and OLAP	•	
ANSI 2003 SQL compliance is required	•	
Preprocessing or exploration of raw unstructured data		•
Online archives alternative to tape		•
High-quality cleansed and consistent data	•	
100s to 1000s of concurrent users	•	•*
Discover unknown relationships in the data	•	•
Parallel complex process logic		•
CPU intense analysis	•	•
System, users, and data governance	•	
Many flexible programming languages running in parallel		•
Unrestricted, ungoverned sand box explorations		•
Analysis of provisional data		•
Extensive security and regulatory compliance	•	
Real time data loading and 1 second tactical queries	•	•*

Figure 8: Comparison between data ware house and hadoop [8]

2.5 Comparing Hadoop and Crossbow

Both Hadoop and crossbow are used to handle or maintain the big data and both are software tool. Both uses map reduce method to handle the big data. So both performances is good. While crossbow uses pipelined method to achieve parallelism so its performance much better in terms of speed.

2.6 Limitation of Hadoop/Map Reduce [11]

Hadoop cannot controls the order in which the maps and the reductions steps run. It is a stateless process. Indexed databases are faster in execution then a Map Reduced work on un indexed data. Reduce step cannot be stopped until all Maps have completed.

3. Example of Human Genome using Big data

Example 1: Cancer Knowledge Action Network [12] use cloud database of Cancer genome help in making sense of Cancer Genomic data with the help of Hadoop. Using Hadoop scientists extracts useful patterns of genome which helps to recognize the mutation of cancer cells and find out the way to cure Cancer.

Example 2: With Human Genome and HapMap Project [13], it is possible to explore subtle genetic influences and many other diseases like diabetes, asthma, migraine, schizophrenia and many more. The figure [14] given below will categories the prevalence as well as the genes or chromosomes associated with human genetic disorders.

Disorder	Prevalence	Chromosome or gene involved
Chromosomal conditions		
Down syndrome	1:600	Chromosome 21
Klinefelter syndrome	1:500–1000 males	Additional X chromosome
Turner syndrome	1:2000 females	Loss of X chromosome
Sickle cell anemia	1 in 50 births in parts of Africa; rarer elsewhere[54]	β-globin
Cancers		
Breast/Ovarian cancer (susceptibility)	~5% of cases of these cancer types	BRCA1, BRCA2
FAP (hereditary nonpolyposis coli)	1:3500	APC
Lynch syndrome	5–10% of all cases of bowel cancer	MLH1, MSH2, MSH6, PMS2
Neurological conditions		
Huntington disease	1:20000	Huntingtin
Alzheimer disease - early onset	1:2500	PS1, PS2, APP
Other conditions		
Cystic fibrosis	1:2500	CFTR
Duchenne muscular dystrophy	1:3500 boys	Dystrophin

Figure 9: Relation among disorders and chromosome [14]

Example3: Human Genome Project [15] is the first completed project related to human genome which helps to read nature's complete genetic blueprint for making a human being. By using the useful information provided by Human Genome Project many companies have launched multiple programs which are dedicated to find out the responsible genes for these diseases. The following figure shows time to time gene discoveries for common complex disease in the field of human genome

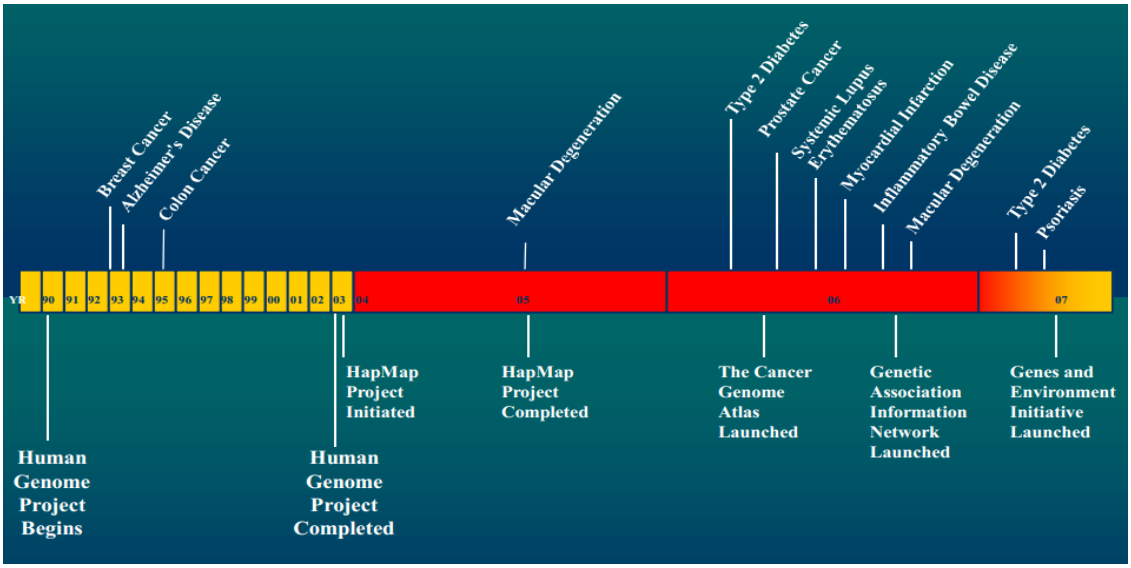

Figure 10: New discovered genes for complex diseases [15]

Example4: For sequencing DNA and analysis of genome sequencing many tools are used which helps to find out disorder in gene and helps to get the similarities in species. For example researchers have found that two-thirds of human genes known to be involved in cancer have counterparts in the fruit fly. A March 2000 study [16] comparing the fruit fly genome with the human genome discovered that about 60 percent of genes are conserved between fly and human. The tools are Human Genome Analysis Tool Kit, Crossbow, Atlas, Sequence Similarity Search, and many more.

As we seen the entire example are related to life of animals or humans. If the result or conclusion or observations are made over large sample using big data then its conclusion will be much more reliable.

4. Conclusions

Whatever I learn from hadoop, it is very good to handle big data. But it is very difficult to implement it. So we can use PIG (i.e. extension of Hadoop) or Hive. Hadoop increases the performance and reliability of data. But Crossbow has much more speed of computation with respect to Hadoop due to pipe line. At last we can say that bio research is related to life of human or animals so we must use big data to analyze large sample for research purpose in place of traditional relational data base or data ware housing. Using big data tools like Genome Analysis tool kit and Crossbow, to decipher DNA is possible. These tools help in finding next generation sequencing of genomes with less costing and in time. By doing sequencing of genomes researchers are able to get knowledge to determine boundaries for DNA Protein interactions and microbial diversity in human and in environment. In depth of DNA the greatest mystery of life is

hidden that how a complete body is formed by a fertilized egg. So, we must analyze human genome with the help of big data and tool should be developed to match or analyze genome of animal or human.

5. Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Mr. Shiv Kumar for the continuous support of my study, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this paper. I could not have imagined having a better advisor and mentor as a student. Besides my advisor, I would like to thank the rest of my Department professors or lecturers or students: Assistant professor B.L. pal, Amit Bhati and Rohit Maheswari, for their encouragement, insightful comments, and hard questions. Last but not the least; I would like to thank our Shiv sir best friend, Sneha Rani supporting me spiritually throughout my life.

References

[1] Hadoop Overview. http://wiki.apache.org/ hadoop/ ProjectDescription
[2] Merina Maharjan " Genome Analysis with MapReduce" June 15, 2011,pp 3-4
[3] An oracle White Paper January 2010,"Integrating Hadoop Data with Oracle Parallel Processing
[4] Michael C. Schatz, Ben Langmead1, Jimmy Lin, Mihai Pop, Steven L. Salzberg "Whole Genome Resequencing Analysis in the Clouds"
[5] http://www.tutorialspoint.com /dwh/dwh_architecture.html
[6] http://cbse.soe.ucsc.edu/research/human_genome
[7] Michael C. Schatz, Ben Langmead1, Jimmy Lin, Mihai Pop, Steven L. Salzberg "Whole Genome Resequencing Analysis in the Clouds"
[8] Casey McTaggart ,"Object-oriented framework presentation", Hadoop/MapReduce
[9] M.C. Srivas "talks about the MapR architecture", "http//:www.Hadoop Architecture Matters MapR.htm"
[10] Dr.amr Awadallah,"Hadoop and the data warehouse:,when to whic use", pp18 &19
[11] http://www.cs.colorado.edu/~kena/classes/5448/s11/presentations/hadoop.pdf
[12] http://www.eweek.com/c/a/Health-Care-IT/Verizon-NantWorks-Using-Big-Data-for-Cancer-Treatment-587749/
[13] http://en.wikipedia.org/wiki/International_HapMap_Project
[14] http://en.wikipedia.org/wiki/Human_genome
[15] http://www.genome.gov/10001772
[16] http://www.genome.gov/11006946

Author Profile

Poonam Kumari is currently pursuing masters degree program in Computer science and engineering in Mewar University, India.

Shiv Kumar received the M. Tech. degree in Computer Science and Engineering from Mewar University Chittorgargh in 2012. During 2007-2013, he stayed in Canon India Private limited Center of Excellence center and India Software Center Noida and Gurgaon of India. He know with Mewar University, Chittorgargh.