**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

[www.msrit.edu](http://www.msrit.edu)

**BANGALORE-560054**



**Big Data Analysis Using SP Theory Of Intelligence**

**Submitted by**

Chaithra B S 1MS13CS403

Mala S 1MS13CS411

Varsha C G 1MS13CS422

Gaurav 1MS12CS142

**Guided by**

Soumyarani C N

Professor

Department of Computer Science and Engineering

**CONTENTS**

**1 SYNOPSIS**

Overview

Problem statement

Existing System

Proposed System

Application

Expected Outcomes

**2 PROJECT PLAN**

Effort estimation

Gantchart

Risk Identification

**3 SOFTWARE REQUIREMENT SPECIFICATIONS**

Functional Requirements

Non-Functional Requirements

System Requirements

External interface

User interface

**4 LITERATURE SURVEY**

**5 DESIGN DOCUMENT TEMPLATES**

Use case diagrams

System architecture

Data flow diagrams

**6 REFERENCES**

1. **INTRODUCTION**

**OVERVIEW**

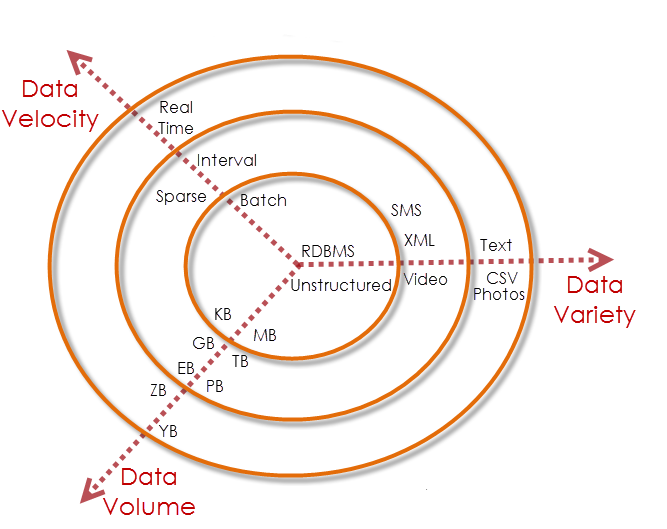
Big data the large volumes of data that are now produced in many fields can present problems in storage, transmission, and processing, but their analysis may yield useful information and useful insights. In broad terms, the potential benefits of the SP system, as applied to big data, are in these areas.

* Overcoming the problem of variety in big data. Harmonizing diverse kinds of knowledge, diverse formats for knowledge, and their diverse modes of processing, via a universal framework for the representation and processing of knowledge.
* Learning and discovery. The unsupervised learning or discovery of ‘natural’ structures in data.
* Interpretation of data. The SP system has strengths in areas such as pattern recognition, information retrieval, parsing and production of natural language, translation from one representation to another, several kinds of reasoning, planning and problem solving.
* Velocity: analysis of streaming data. The SP system lends itself to an incremental style, assimilating information as it is received, much as people do.

**PROBLEM STATEMENT**

The volume, velocity and variety are the challenging problems of big data. The large amount of data transferred from different sources at rapid speeds need to be stored efficiently. The simple and power theory of intelligence helps in tackling the challenging problems of big data and enhances the efficiency of data storage and data retrieval.

**DEFINING BIG DATA**

The 3Vs that define *Big Data* are Variety, Velocity and Volume.

**Volume**

We currently see the exponential growth in the data storage as the data is now more than text data. We can find data in the format of videos, musics and large images on our social media channels. It is very common to have Terabytes and Petabytes of the storage system for enterprises. As the database grows the applications and architecture built to support the data needs to be reevaluated quite often. Sometimes the same data is re-evaluated with multiple angles and even though the original data is the same the new found intelligence creates explosion of the data. The big volume indeed represents *Big Data*.

**Velocity**

The data growth and social media explosion have changed how we look at the data. There was a time when we used to believe that data of yesterday is recent. The matter of the fact newspapers is still following that logic. However, news channels and radios have changed how fast we receive the news. Today, people reply on social media to update them with the latest happening. On social media sometimes a few seconds old messages (a tweet, status updates etc.) is not something interests users. They often discard old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represent *Big Data*.

**Variety**

Data can be stored in multiple format. For example database, excel, csv, access or for the matter of the fact, it can be stored in a simple text file. Sometimes the data is not even in the traditional format as we assume, it may be in the form of video, SMS, pdf or something we might have not thought about it. It is the need of the organization to arrange it and make it meaningful. It will be easy to do so if we have data in the same format, however it is not the case most of the time. The real world has data in many different formats and that is the challenge we need to overcome with the *Big Data*. This variety of the data represent represents *Big Data*.

**EXISTING SYSTEM**

In the big data platform we faced many problems like storing the data, managing the data and speed of data processing.

**Disadvantages of existing system**

* Problem of velocity in big data
* Problem of storing and managing big data
* Low speed of data processing

**PROPOSED SYSTEM**

In this paper, we proposed Simplicity and Power theory of intelligence to overcome the variety of problems in big data platform.

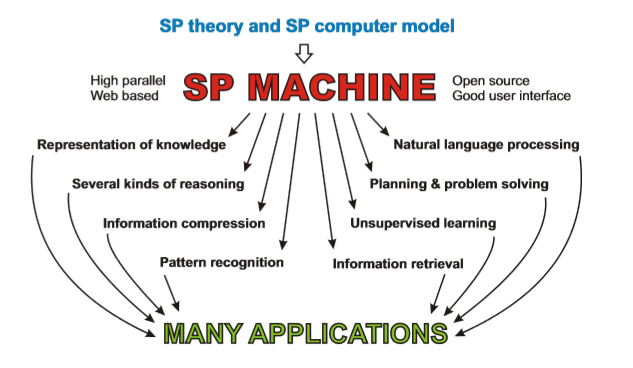
**Advantages of proposed system**

* Volume: making big data smaller
* Velocity: analysis of streaming data.
* The unsupervised learning or discovery of ‘natural’ structures in data

**APPLICATION OF THE PROJECT**

**The SP theory has three main elements:**

* All kinds of knowledge are represented with patterns: arrays of atomic symbols in one or two dimensions.
* At the heart of the system is compression of information via the matching and uniﬁcation (merging) of patterns, and the building of multiple alignments.
* The system learns by compressing “New” patterns to create “Old” patterns.



The SP theory of intelligence combines conceptual simplicity with descriptive and explanatory power in several areas, including concepts of “computing”, the representation of knowledge, natural language processing, pattern recognition, several kinds of reasoning, the storage and retrieval of information, planning and problem solving, unsupervised learning, information compression, neuroscience, and human perception and cognition. In the SP machine there is potential for the simplification of computing systems, including software, with corresponding savings in the time, effort and cost in the development of applications, and other beneﬁts related to information compression.

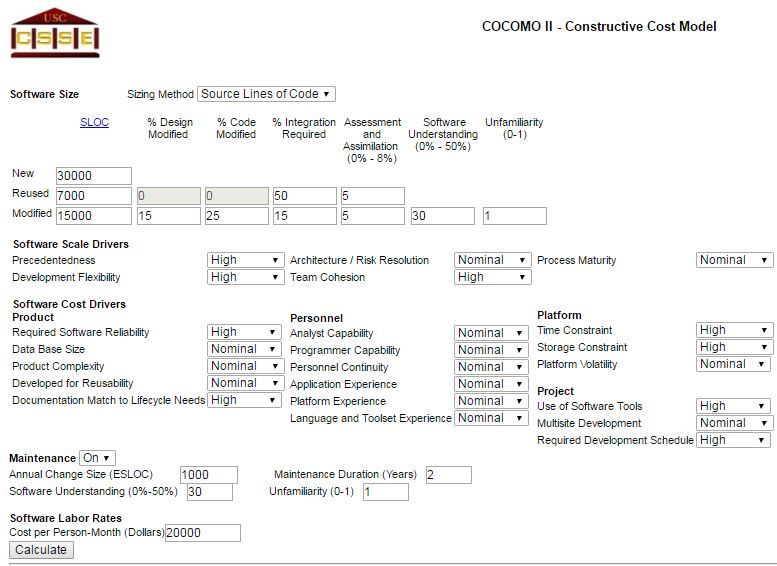
**EXPECTED OUTCOMES**

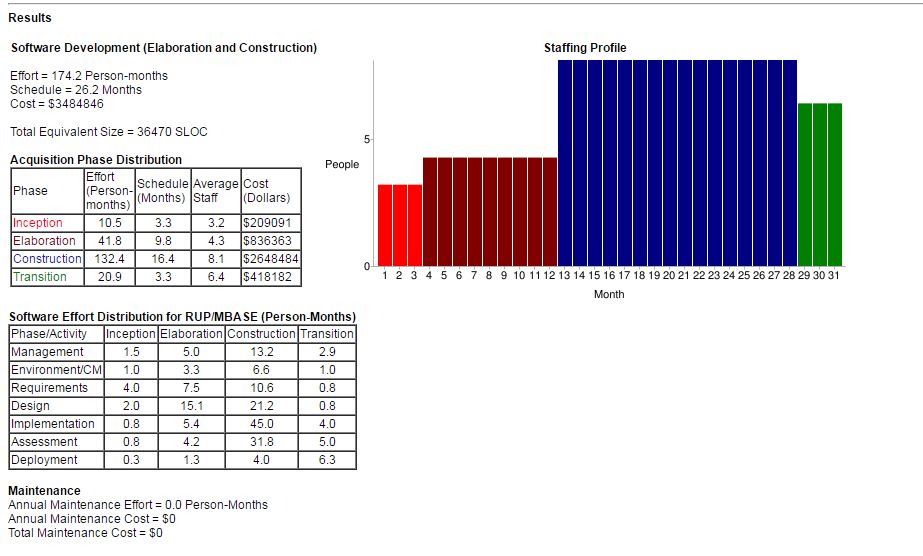
* Making big data smaller.
* Analysis of streaming data.
* The unsupervised learning or discovery of ‘natural’ structures in data.

1. **PROJECT PLAN**

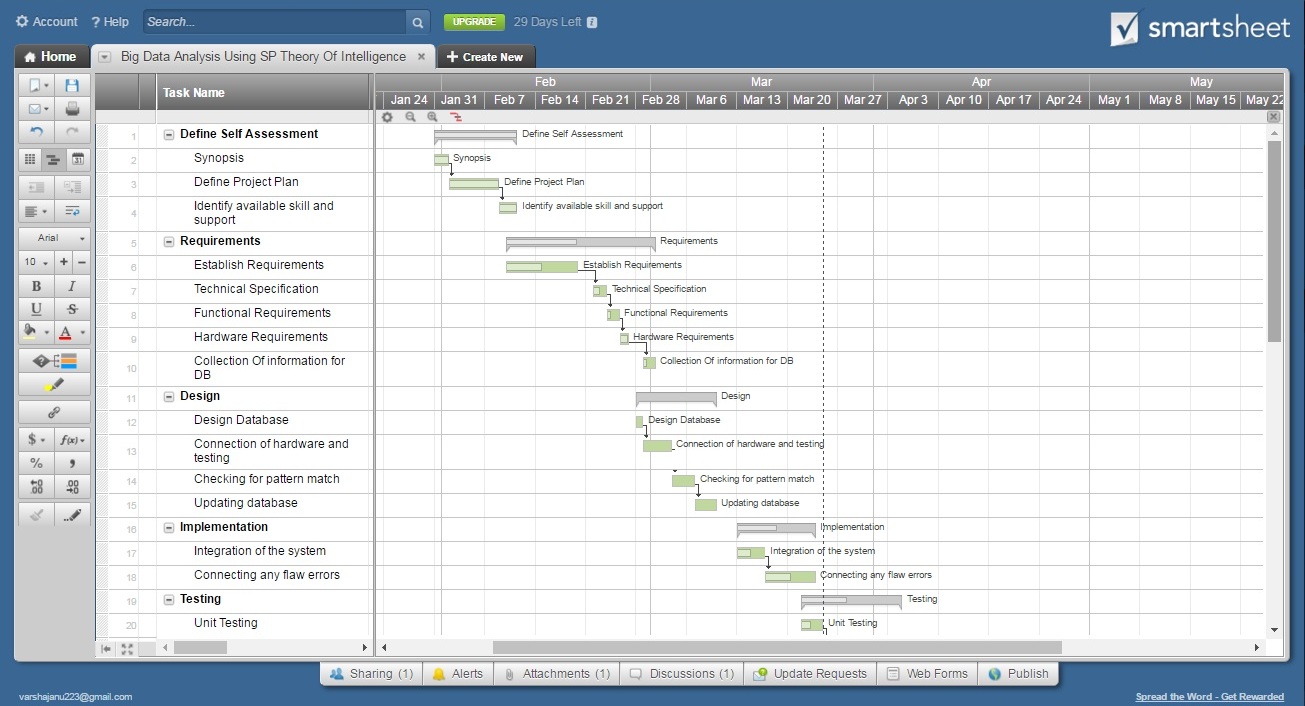
Effort Estimation using COCOMO Model

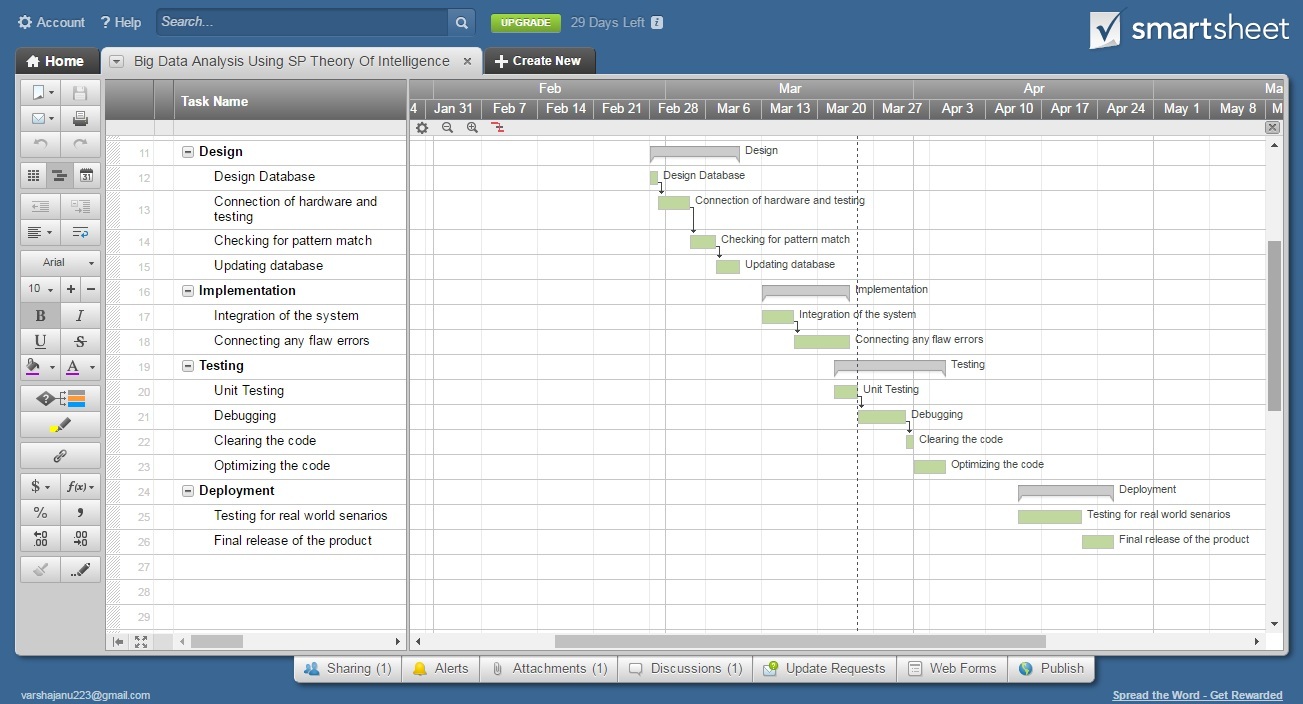
Top of Form

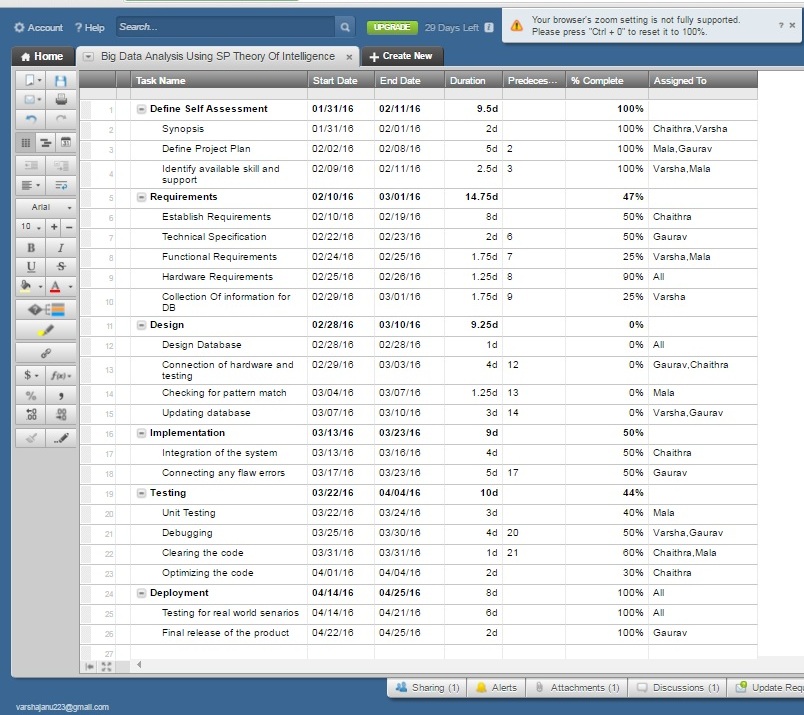


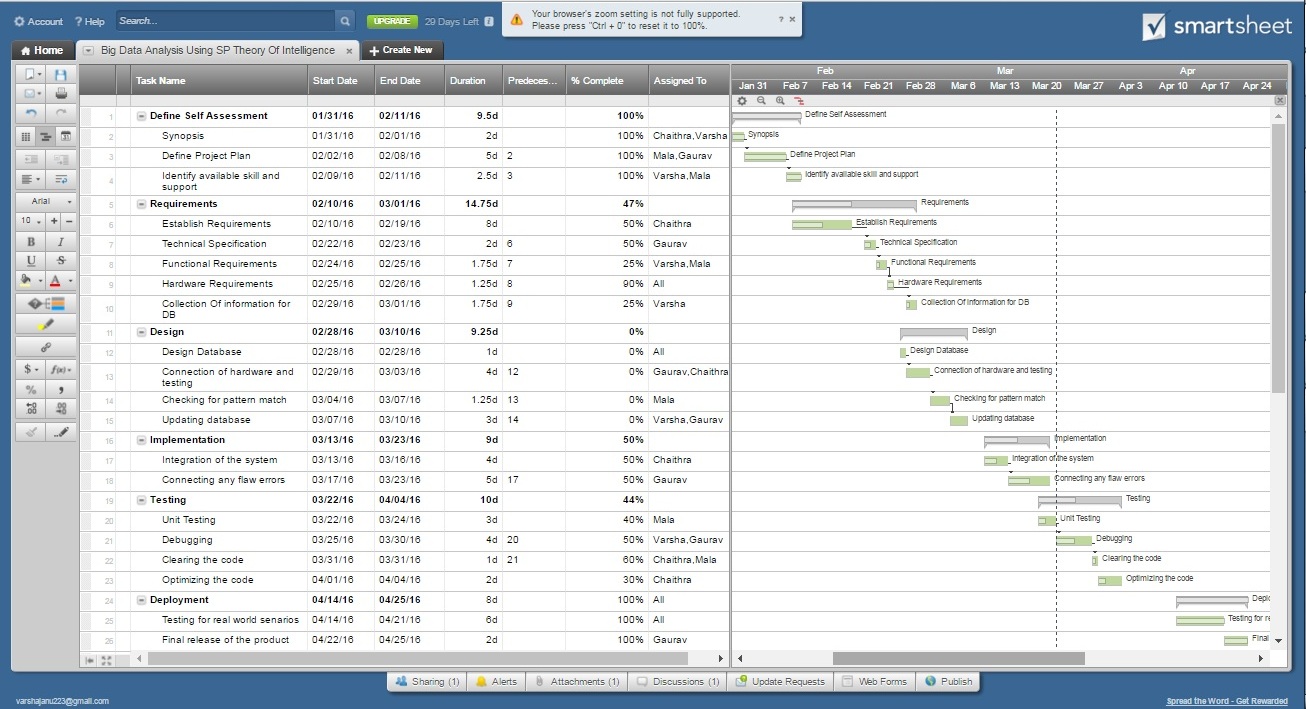


Schedule Using Gantt Chart









Risk Identification (Top 5 Risk) with Risk Mitigation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Risk No** | **Risk** | **Event Type** | **Impact** | **Probability Of Occurrence** | **Risk Rate** | **Mitigation Steps** |
| 1 | Personnel Shortfalls | Cost | Severe | 60% | High | * Training on the programming * Hiring Experts |
| 2 | Unrealistic Schedules | Schedule | Critical | 75% | High | * Consult the guide about the problem * Proper Planning is required |
| 3 | Developing Wrong Software Function | Performance | Critical | 70% | Medium | * Understanding the SRS * Proper documentation of requirements |
| 4 | Updating Database Error | Performance | Severe | 60% | Medium | * Usage of well developed software * Proper testing before giving to the user |
| 5 | Inexperience in the problem domain | Cost,  Schedule  ,Performance | Severe | 70% | High | * Experienced developers must be used for development of the given software * Proper survey problem domain |

1. **SYSTEM REQURIEMENTS AND SPECIFICATION**

**HARDWARE REQUIREMENTS**

Processor : Pentium IV 2.4 GHz

Hard Disk : 250 GB

RAM : 1 GB

**SOFTWARE REQUIREMENTS**

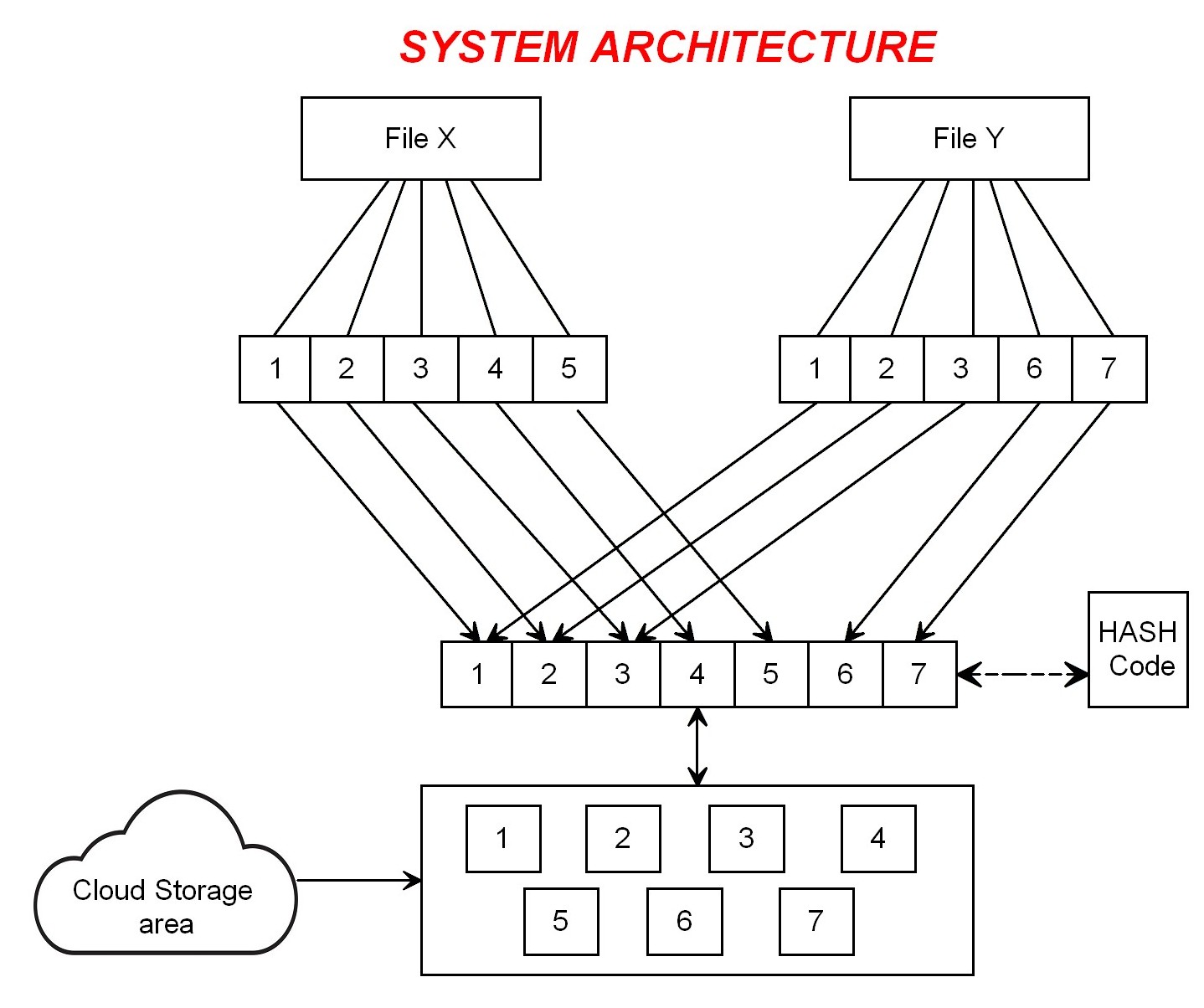
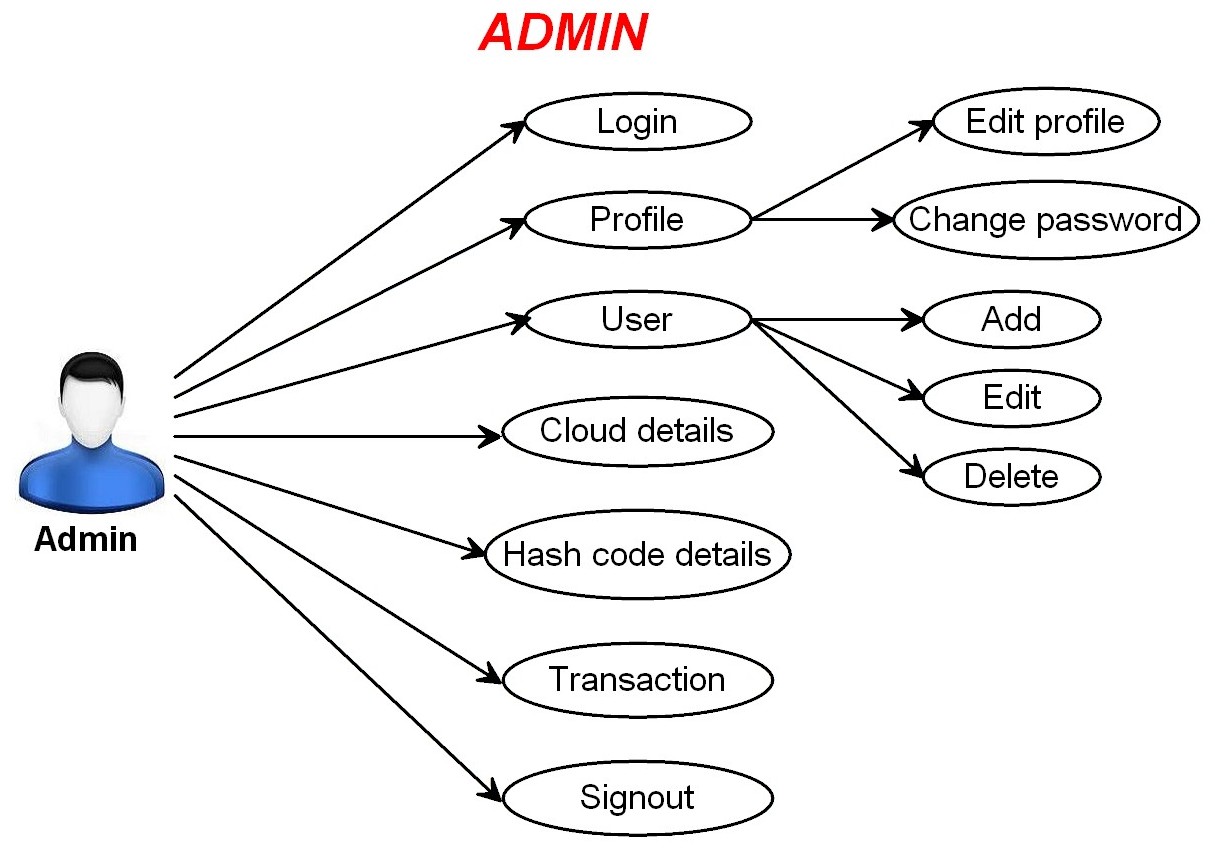
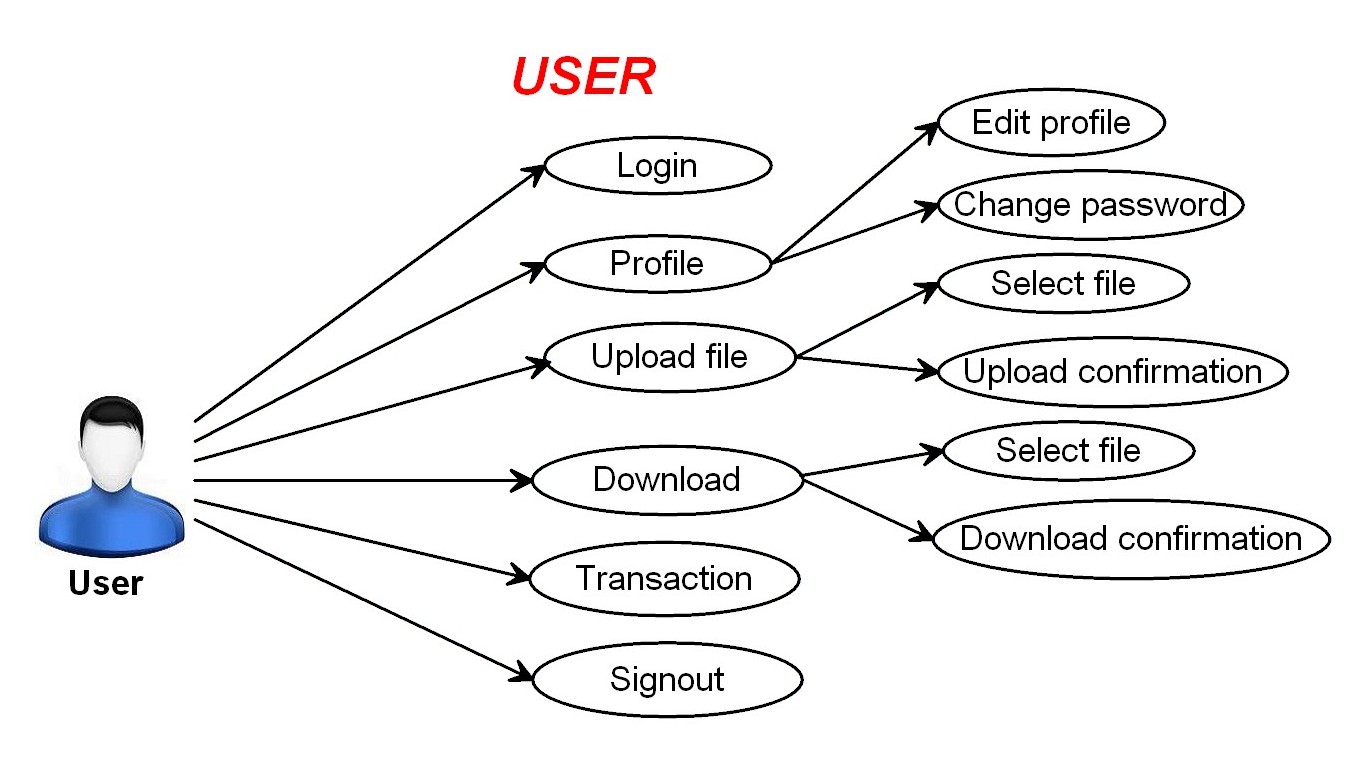
Operating system : Microsoft windows

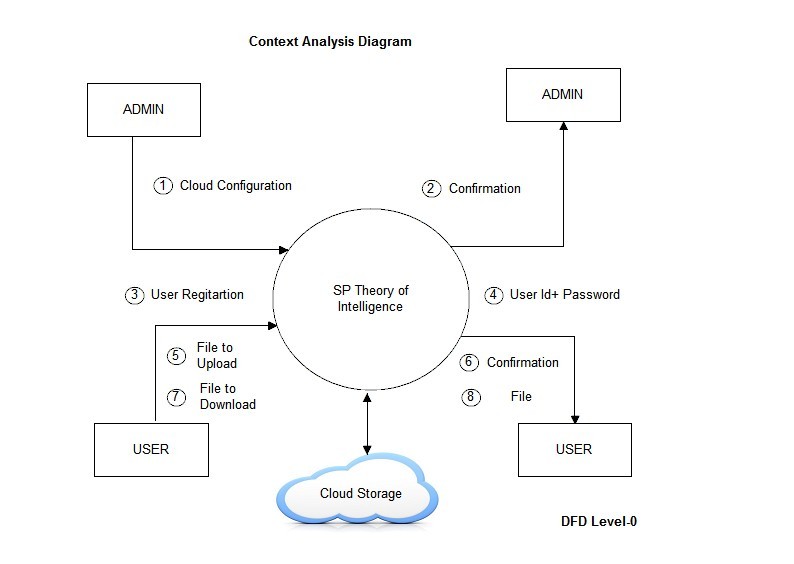
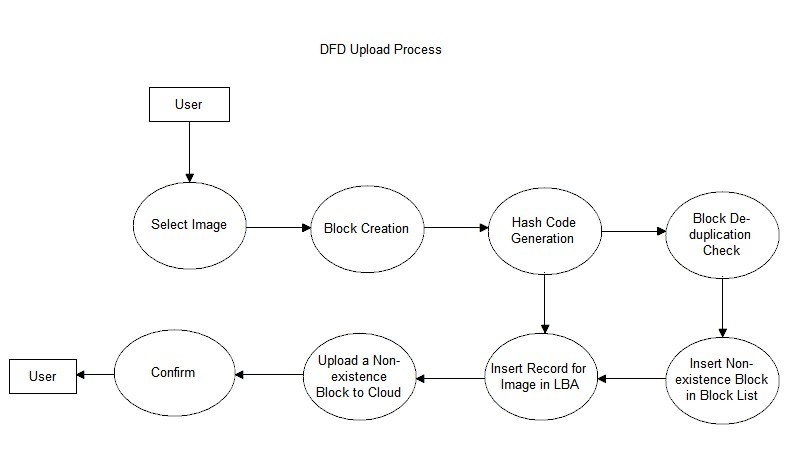
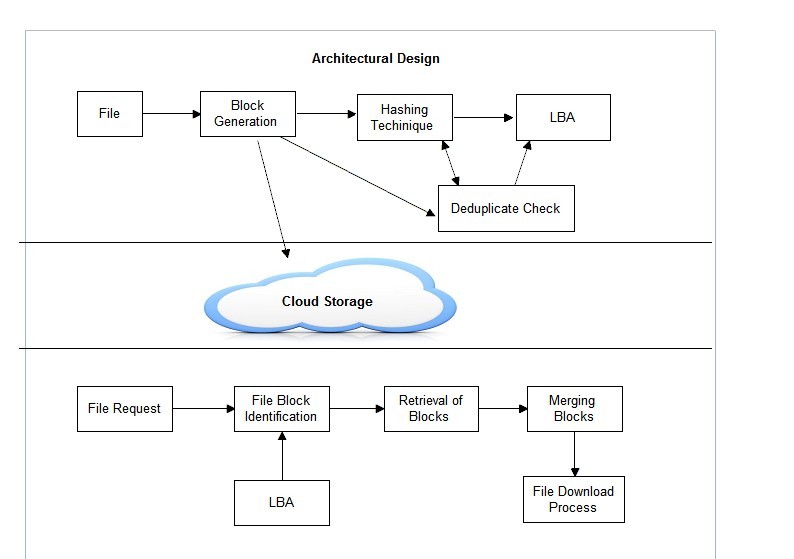
Coding Language : Java (JDK 1.7)

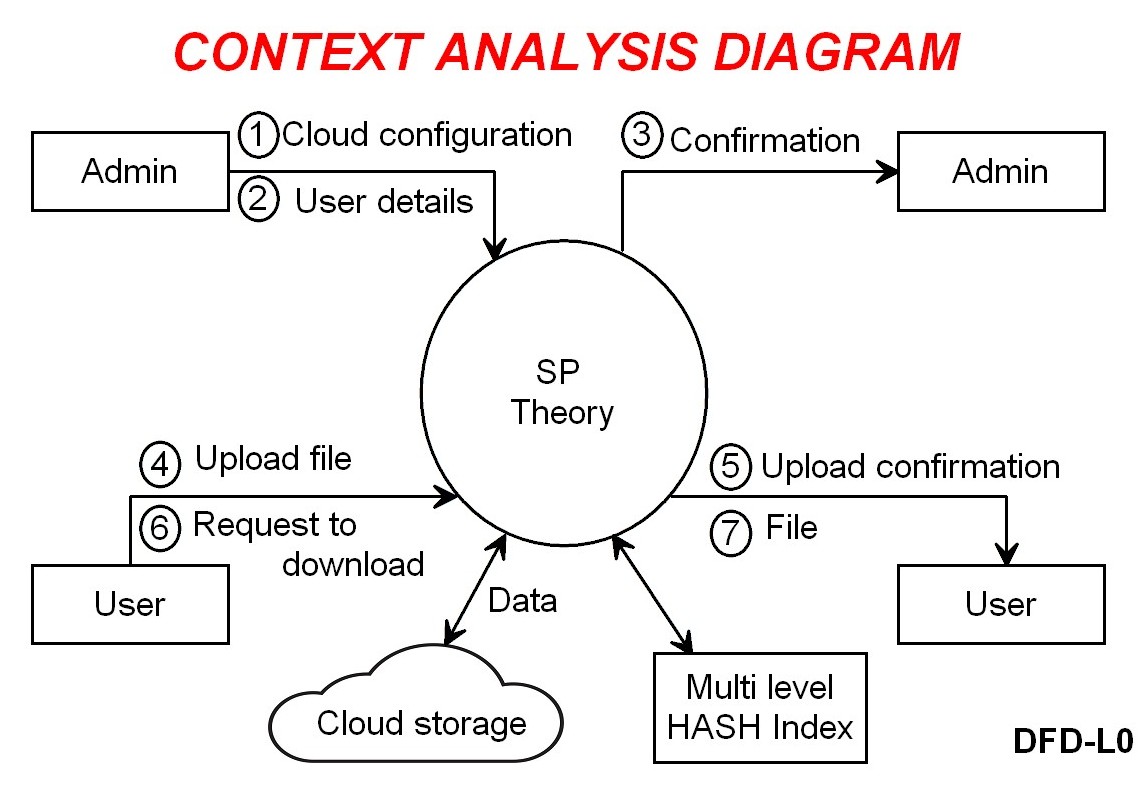
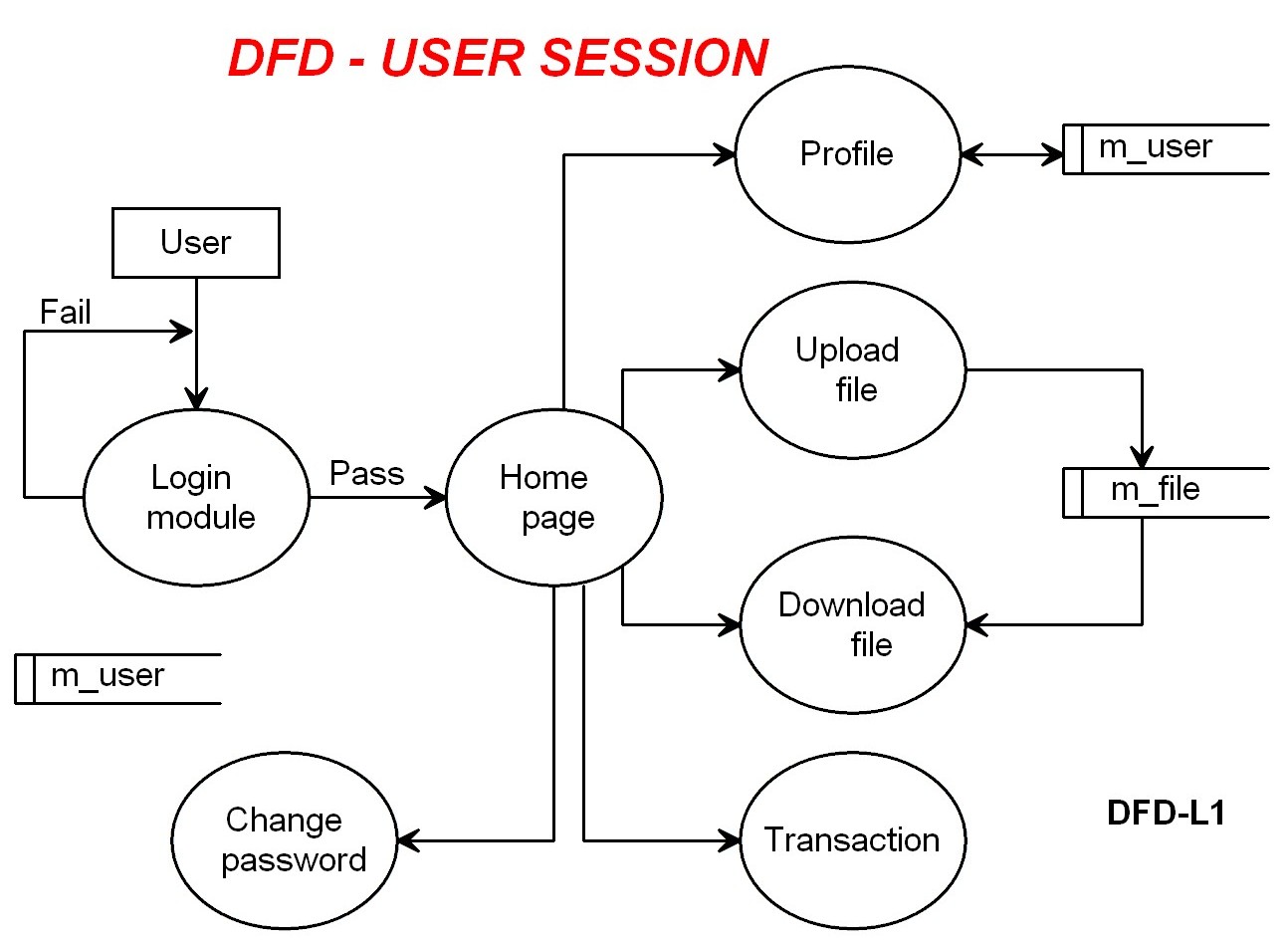
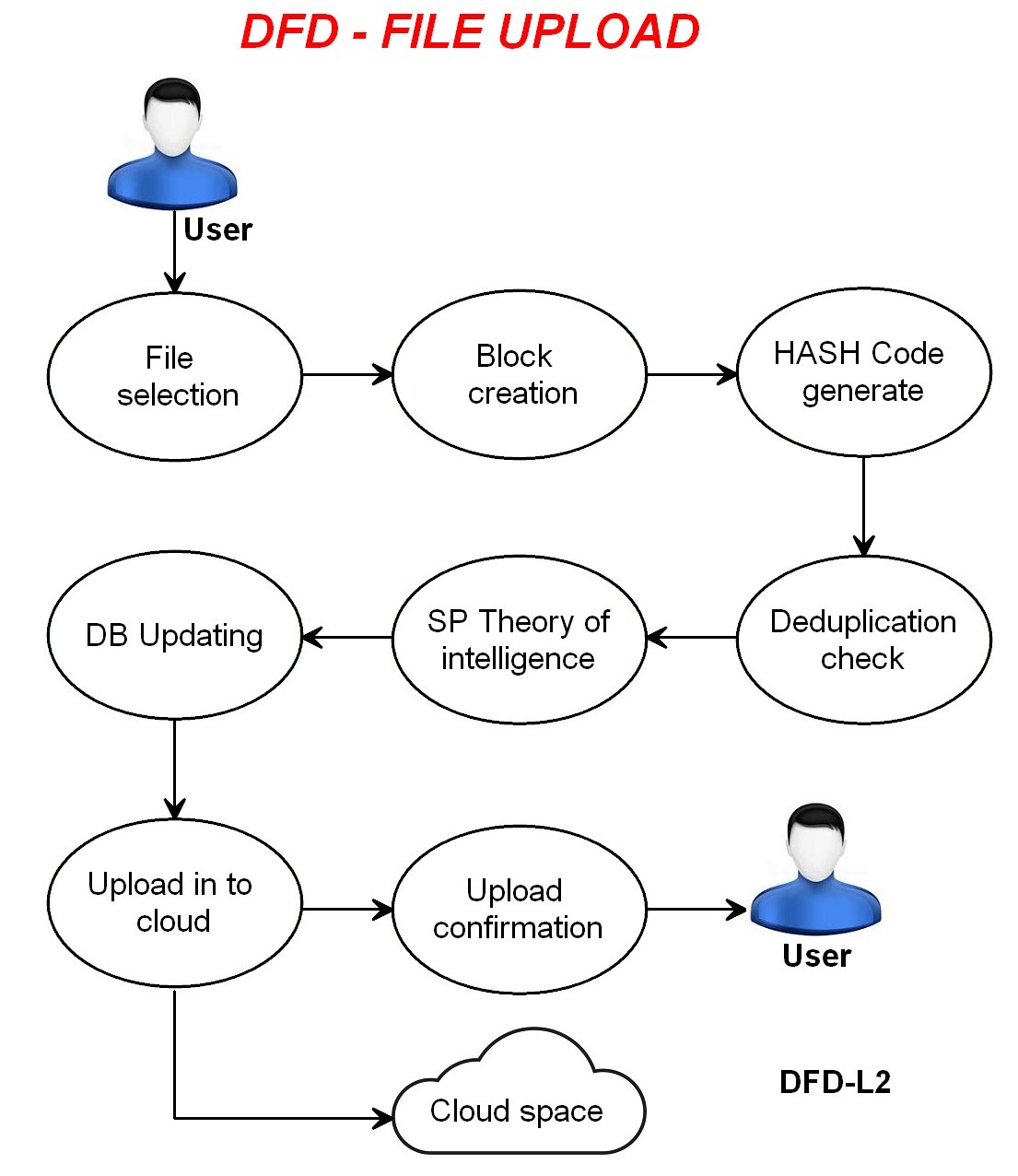
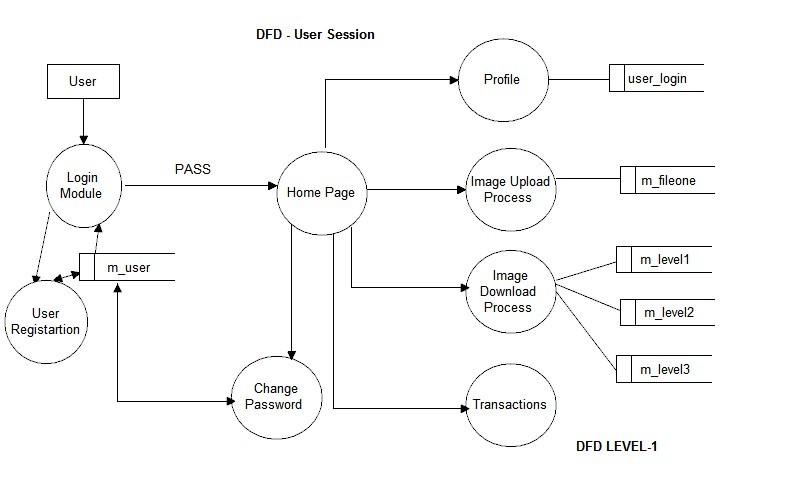
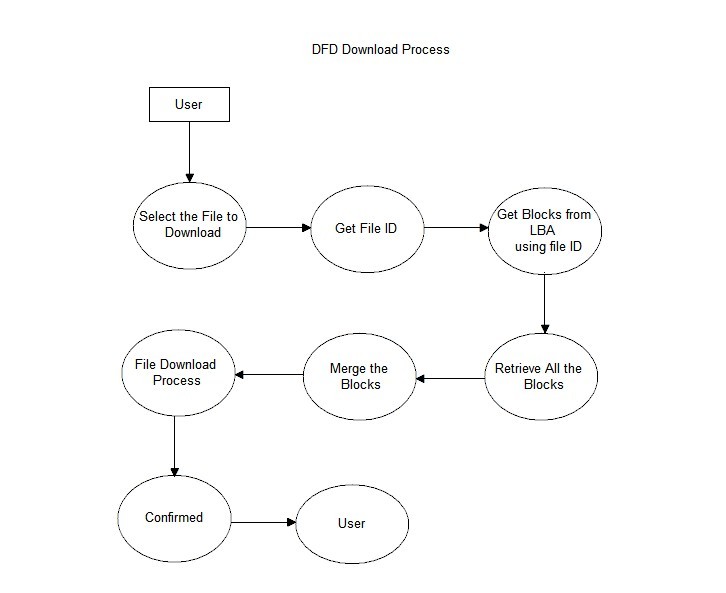
Eclipse tool : Eclipse Indigo

Database : My-SQL 5.0

Database GUI : SQLYog

1. **DESIGN DOCUMENT TEMPLATE**





**REFERENCES**

[1] J. G. Wolff, “The SP theory of intelligence: An overview”, “Information compression, intelligence, computing and mathematics”, vol. 4, no. 3, pp. 283 341, 2013.

[2] J. G. Wolff, National Research Council, “Frontiers in Massive Data Analysis”. Washington, DC, USA: National Academies Press, 2013.

[3] J. G. Wolff, “The SP theory of intelligence: Benefits and applications”, “Information compression, intelligence, computing and mathematics’’, vol. 5, no. 1, pp. 1 27, 2014

[4] J.G.Wolff, ‘‘Application of the SP theory of intelligence to the under-standing of natural vision and the development of computer vision’’ 2013, in preparations