BERSAMA-SAMA MELAKSANAKAN TRANSFORMASI

METHODOLOGY

Analitis Data Raya Sektor Awam (DRSA)

DRSA Analitis Data Raya Sektor Awam

TABLE OF CONTENTS

TITLE **PAGE**

1. INTRODUCTION 1
1.1. PURPOSE 1
1.2. SCOPE 1
2. INTRODUCTION TO KEY CONCEPTS 2
2.1. BIG DATA 2
2.2. DATA SCIENCE, DATA ENGINEERING AND DATA SCIENCE PROCESS 5
2.3. CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING 7
3. NIST BIG DATA INTEROPERABILITY FRAMEWORK 12
4. DRSA BIG DATA METHODOLOGY 13

LIST OF ILLUSTRATIONS

TITLE **PAGE**

Illustration 1: Data science and data engineering 5
Illustration 2: Harvard CS109 Data Science Process 6
Illustration 3: Phases in CRISP-DM methodology 9
Illustration 4: Stages in Malaysian Public Sector Big Data Analytic Methodology 13
Illustration 5: Stages 1 – Business Understanding 14
Illustration 6: Stages 2 – Requirements Definition 15
Illustration 7: Stages 3 – Business Understanding 16
Illustration 8: Stages 4 – Analytical Model Development 18
Illustration 9: Stages 5 – Data Product Development 19
Illustration 10: Stages 6 – Transition to Production 20
Illustration 11: Stages 7 – Monitoring 21

DRSA METHODOLOGY

1. Introduction

1.1. Purpose

The purpose of this document is to provide a methodology and process framework for implementation of a Big Data project. The methodologies and processes defined in this document was developed based on lesson learned throughout the implementation of Projek Analitis Data Raya Sektor Awam (DRSA) that have been customized to fit the existing processes and practices in MAMPU and government agencies in general.

1.2. Scope

This document provides a high level understanding of key concepts in Big Data, iterative development, and processes that can be followed for future project, which includes the following topics:

i. Preparation and planning for a Big Data project

ii. Stages for a Big Data project and high level implementation tasks

iii. Setting up a data science or data analytic team

iv. Documents and deliverables that shall be created throughout the project

v. Managing development progress

vi. Transitioning to production

vii. Post-development monitoring of the project

2. Introduction to Key Concepts

This section provide high level understanding on key concepts in the ecosystem of Big Data to assist in understanding the base concepts, methodologies and standards used for the development of this methodology, which are:

i. Harvard CS109: Data Science Process – Provides high level process flow for execution of explorative analytics and data science which is the foundation of this methodology.

ii. Cross Industry Standard Practice for Data Mining (CRISP-DM) – Defines high level process flow for project delivery of machine learning and data mining based projects. CRISP-DM provides the majority of core processes and tasks for this methodology.

iii. National Institute of Standards and Technology Big Data Interoperability Framework (NIST-BDIF) – Defines standards for defining and documenting Big Data use-cases and requirements.

2.1. Big Data

> *"Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data*[1]
>
> - McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity, May 2011

1 http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

> *Big data analytics refers to the strategy of analyzing large volumes of data, or big data. This big data is gathered from a wide variety of sources, including social networks, videos, digital images, sensors, and sales transaction records. The aim in analyzing all this data is to uncover patterns and connections that might otherwise be invisible, and that might provide valuable insights about the users who created it. Through this insight, businesses may be able to gain an edge over their rivals and make superior business decisions.*[2]
>
> - Techopedia

At the point of time which this document was written, there are no rigorous term that defines Big Data. Different organizations define it differently, but in general a Big Data implementation consist of analyzing data which contains one or many of the following attributes:

i. **Volume** – the analysis of massive amount (terabytes, petabytes) of data to extract insights

ii. **Velocity** – the real-time analysis of streaming data that comes at a very fast rate for faster decision making.

iii. **Variety** – the analysis of data coming from various sources; structured, unstructured, data coming from internal sources of an organization and data coming from external sources.

iv. **Veracity** – the analysis of highly inconsistent and poor quality datasets

2 http://www.techopedia.com/definition/27745/big-data

Besides the attributes of the datasets, Big Data practices also bring into the table data analysis techniques which previously were not accessible to mass market consumers due to technological limitations. Big Data Analytics may consist of the following properties:

i. **Descriptive Analytics** – the analysis to summarize what have happened (hindsight), and to analyze why it happens (insight)

ii. **Predictive Analytics** – the analysis to estimate, deduce and forecast (foresight) new information based on recent and historical data. Predictive analysis is not only about providing a forecast of the future , but also to estimate new, unknown property of existing data based on its attributes. In general, it is to take data that you have, to predict data that you do not have.

iii. **Prescriptive Analytics** – this analysis is a type of predictive analytics which prescribe an action, so that the business decision-makers can take this information and act upon. Predictive analytics doesn't predict one possible future, but rather "multiple futures" based on the decision-maker's actions. Predictive analytics requires actionable data and a feedback system that tracks the outcome produced by the action taken.

In order to analyze collected datasets to provide the stated analytics, various Big Data tools and techniques will be used such as:

i. **Statistical analysis** – analysis that involve scrutinizing every data sample to describe the nature of the data and explore the relationships of the data to the rest of available data.

ii. **Machine learning** – an analytic process that involve in the design of pattern recognition algorithms that can learn from , and make prediction s on data

iii. **Data mining** – an analytic process that applies machine learning and other techniques to collect, store, analyze, and extract value from mined datasets.

iv. **Data Warehouse / Data Lake** – a repository where data are stored for the purpose of reporting and analysis. Data warehouse have a different architecture than data lake, with different pro and cons, but they both serve the same purpose.

The end goal of analyzing Big Data is to harness information in novel ways to produce useful insights, and create new forms of value for the business

2.2. Data Science, Data Engineering and Data Science Process

Data science is a field of practice that focuses on the extraction of knowledge from large volumes of data that are structured or unstructured. It is a continuation of the field of data mining and predictive analytics, also known as Knowledge Discovery and Data Mining (KDD)[3]. Data science involves principles, processes and techniques for understanding phenomena via automated analysis of data.

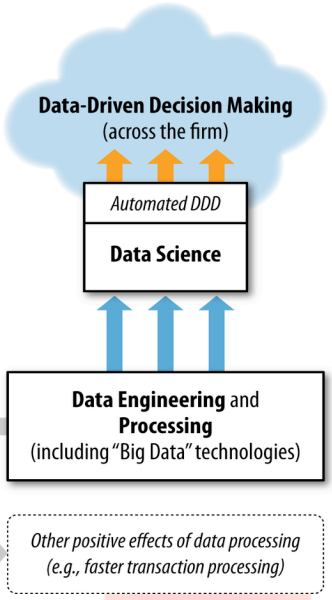

Illustration 1: Data science and data engineering

Data science enables data driven decision making (DDD), which is a practice of

3 https://en.wikipedia.org/wiki/Data_science

basic decision from facts extracted from analyzed data, instead of just purely on intuition[4].

To prepare and provide the data needed for data science activities, data engineers deal with the development of processes, workflow and programs to handle the process of preparation of datasets such as scheduled loading, anonymization, cleansing, and integration.

CS109 Data Science Process is an iterative process framework, defined in Harvard CS109 Data Science course, as a guidance for execution of data science activities.

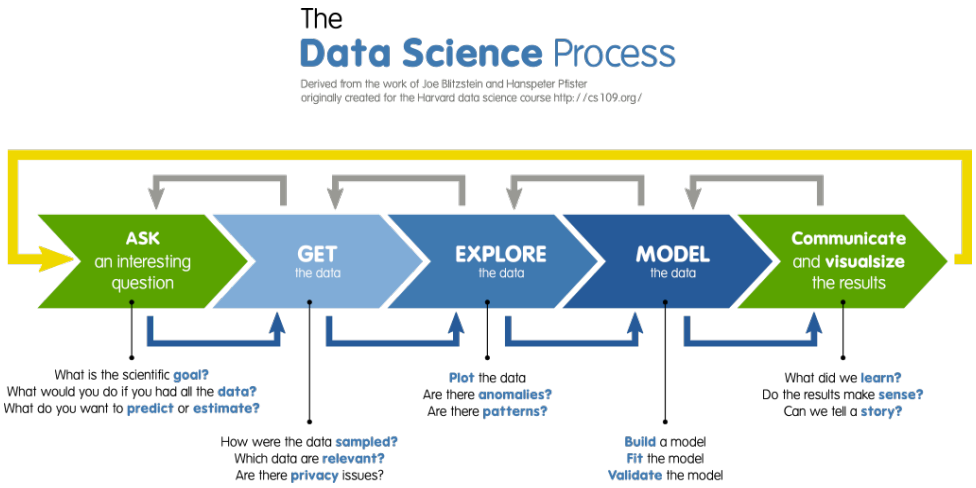

Illustration 2: Harvard CS109 Data Science Process

Data Science Process is organized into 5 iterative stages:

i. Ask an interesting question

ii. Get the data

iii. Explore the data

iv. Model the data

v. Communicate and visualize the results

4 Data Science for Business; Foster Provost & Tom Fawcett; pg 5

In Data Science Process, an analytic activity is started through identifying and **asking a business question** which can provide an impact and benefit to an organization.

Once a business question have been identified, data collection activities are then executed to **gather the necessary data** needed to answer the business question. The data may come from sources such as:

i. Existing datasets in current organization

ii. Acquisition of datasets from external organizations

iii. Data collection exercises through data collection initiatives or projects

iv. Publicly available datasets from internet, social media and Open Data

The next stage in the process is **exploring all gathered datasets** to get to know the data, identify patterns and any anomalies in the data. The information gathered at this stage will be used for development of any models to answer the business question in the subsequent stage.

Once a level of understanding on the data have been reached, and that there are enough data to work with, the data scientist will then engage in the **development of analytical model** utilizing various data analytics techniques and tools such as statistical analysis, machine learning and data mining.

Finally, the results of the analysis are then **communicated and presented through visualization** that can be understood by business stakeholders for assisting in decision making for the organization.

2.3. Cross Industry Standard Process for Data Mining

Cross Industry Standard Process for Data Mining, commonly known by its acronym CRISP-DM is a data mining process model that describes commonly used approaches that data mining experts use to tackle problems. CRISP-DM was

conceived in late 1996 by three “veterans” of the young and immature data mining market. DaimlerChrysler (then Daimler-Benz) was already ahead of most industrial and commercial organizations in applying data mining in its business operations.

CRISP-DM was conceived to provide a standard process model to help in the approach of the implementation for data mining, and address common issues that will be faced by all practitioners.

Many Big Data related development activities, such as predictive analytics and prescriptive analytics are essentially data mining activities, therefore, CRISP-DM have been suggested as one **good reference model for the implementation of Big Data project**[5]. However, CRISP-DM alone is not enough to be a methodology for delivery a full end-to-end Big Data project. It **only covers the development of analytical models**, which is only in the data layer and **does not define any phase for the development of the mechanism to consume the results**, such as analytics dashboards, or data applications.

CRISP-DM breaks the process of data mining into six major phases.[6]

5 Data Science for Business; Foster Provost & Tom Fawcett;
6 https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

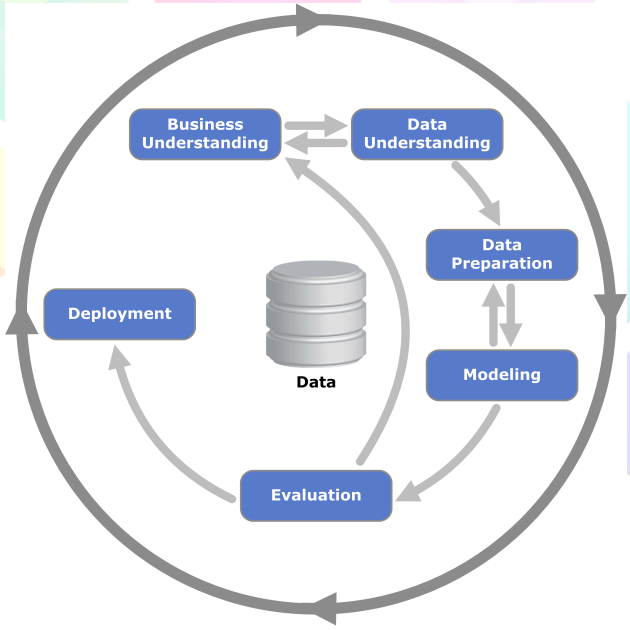

Illustration 3: Phases in CRISP-DM methodology

The sequence of the phases is not strict and moving back and forth between different phases is always required. The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle in the diagram symbolizes the cyclic nature of data mining itself. A data mining process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions and subsequent data mining processes will benefit from the experiences of previous ones.

Phase 1: Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives. A decision model, especially one built using the Decision Model and Notation standard can be used.

Phase 2: Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

Phase 3: Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

Phase 4: Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

Phase 5: Evaluation

At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Phase 6: Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that is useful to the customer. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g. segment allocation) or data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. Even if the analyst deploys the model it is important for the customer to understand up front the actions which will need to be carried out in order to actually make use of the created models.

3. NIST Big Data Interoperability Framework

The National Institute Of Science and Technology (NIST) is leading the development of a Big Data Technology Roadmap that will define and prioritize requirements for interoperability, portability, reusability, and extensibility for big data analytic techniques and technology infrastructure in order to support secure and effective adoption of Big Data. To help develop the ideas in the Big Data Technology Roadmap, NIST is creating the Public Working Group for Big Data.

As part of this initiative, NIST produced a working draft of Big Data Interoperability Framework (NIST-BDIF)[7] which created the following draft standards that help in defining and designing Big Data implementations:

i. Draft SP 1500-1 -- Volume 1: Definitions

ii. Draft SP 1500-2 -- Volume 2: Taxonomies

iii. Draft SP 1500-3 -- Volume 3: Use Case & Requirements

iv. Draft SP 1500-4 -- Volume 4: Security and Privacy

v. Draft SP 1500-5 -- Volume 5: Architectures White Paper Survey

vi. Draft SP 1500-6 -- Volume 6: Reference Architecture

vii. Draft SP 1500-7 -- Volume 7: Standards Roadmap

Volume 3 in NIST-BDIF provides an example in defining Big Data use-cases, functional requirements and non-functional requirements[8], which is used by this methodology as a mechanism to capture the requirements and document into a Business Requirements Specification for the development of Big Data projects. This, in a certain degree, will provide enough information for use-case and requirements interoperability between Malaysian government Big Data project with international organizations that follow the NIST standard.

7 http://bigdatawg.nist.gov/V1_output_docs.php
8 http://bigdatawg.nist.gov/_uploadfiles/M0394_v1_4746659136.pdf

4. DRSA Big Data Methodology

4.1. Overview

DRSA Methodology consist of seven (7) main stages which are further broken down into sub steps and activities. The seven (7) stages are summarized below:

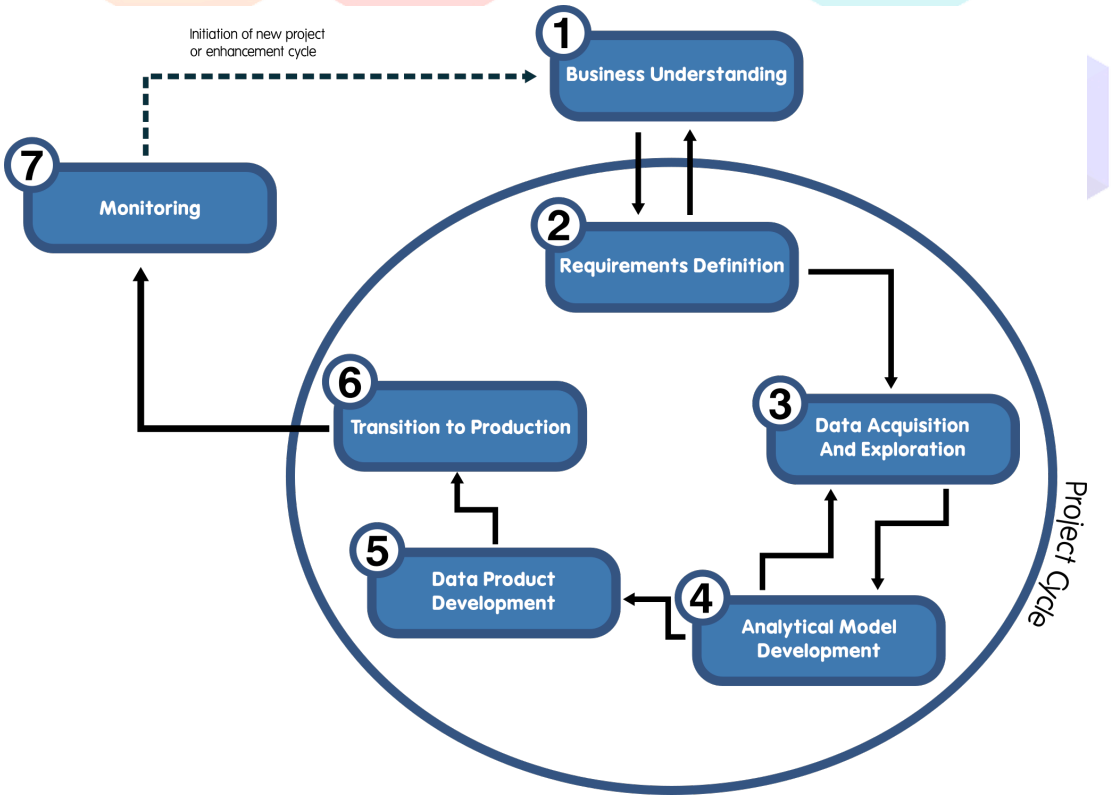

Illustration 4: Stages in Malaysian Public Sector Big Data Analytic Methodology

Stage 1: Busines Understanding

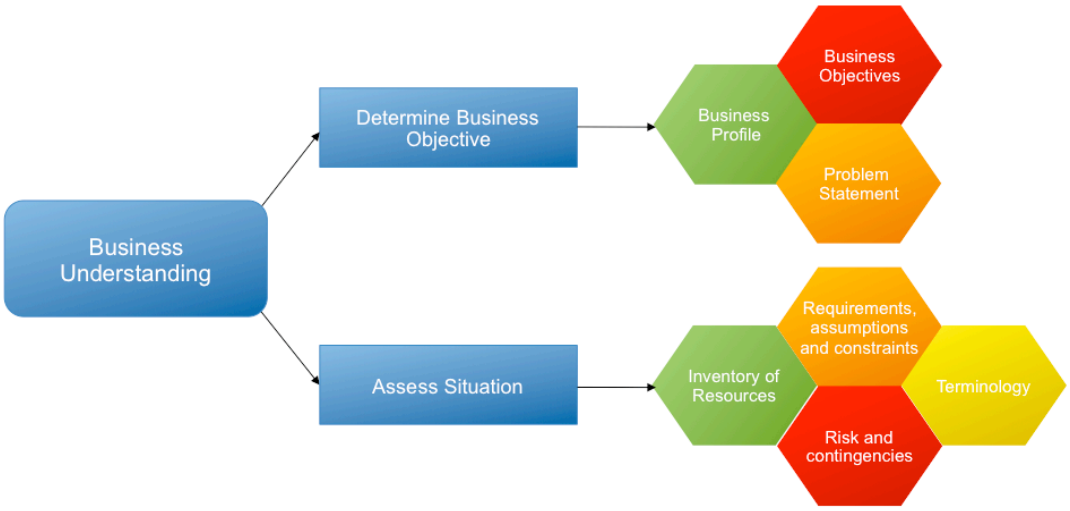

Illustration 5: Stages 1 – Business Understanding

This stage focuses on identifying stakeholders, understanding the business operations and needs, and identifying opportunities from existing and new data that can benefit the business.

Activities and outputs of this stage are:

i. Determine Business Objective
   a. Business Profile
   b. Business objectives
   c. Problem Statement

ii. Assess Situation
   a. Inventory of Resources
   b. Requirements, assumptions and constraints
   c. Risk and contingencies
   d. Terminology

Stage 2: Requirements Definition

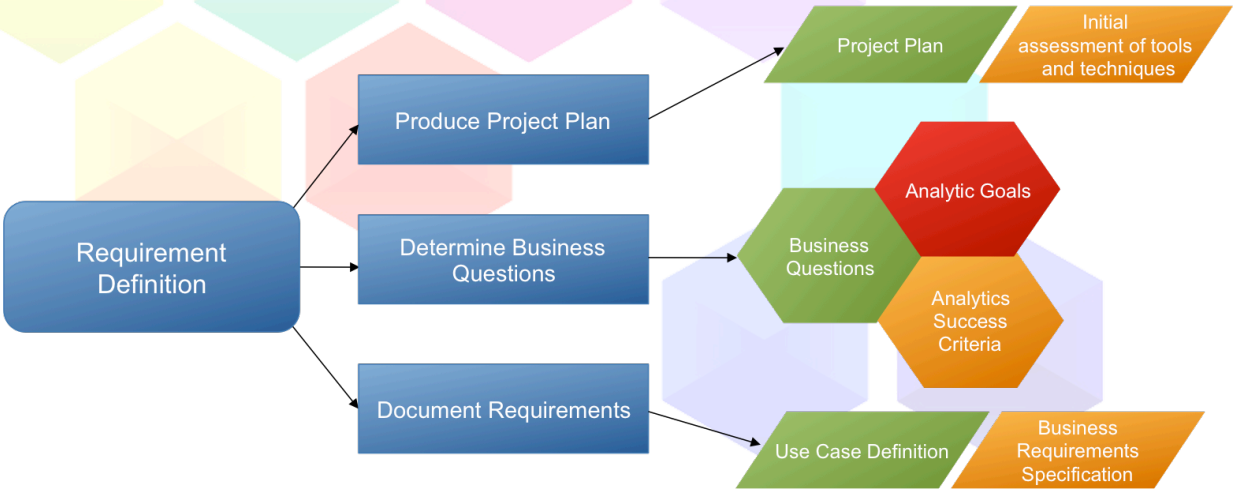

Illustration 6: Stages 2 – Requirements Definition

This stage focuses on defining and documenting the scope of work, business requirements , user requirements and system requirements of the project.

Activities and outputs of this stage are:

i. Produce Project Plan
   a. Project plan
   a. Initial assessment of tools and techniques

ii. Determine Business Questions
   a. Business questions
   b. Analytic goals
   c. Analytic success criteria

iii. Document Requirements
   a. Use case definition
   b. Business Requirements Specification

Stage 3: Data Acquisition and Exploration

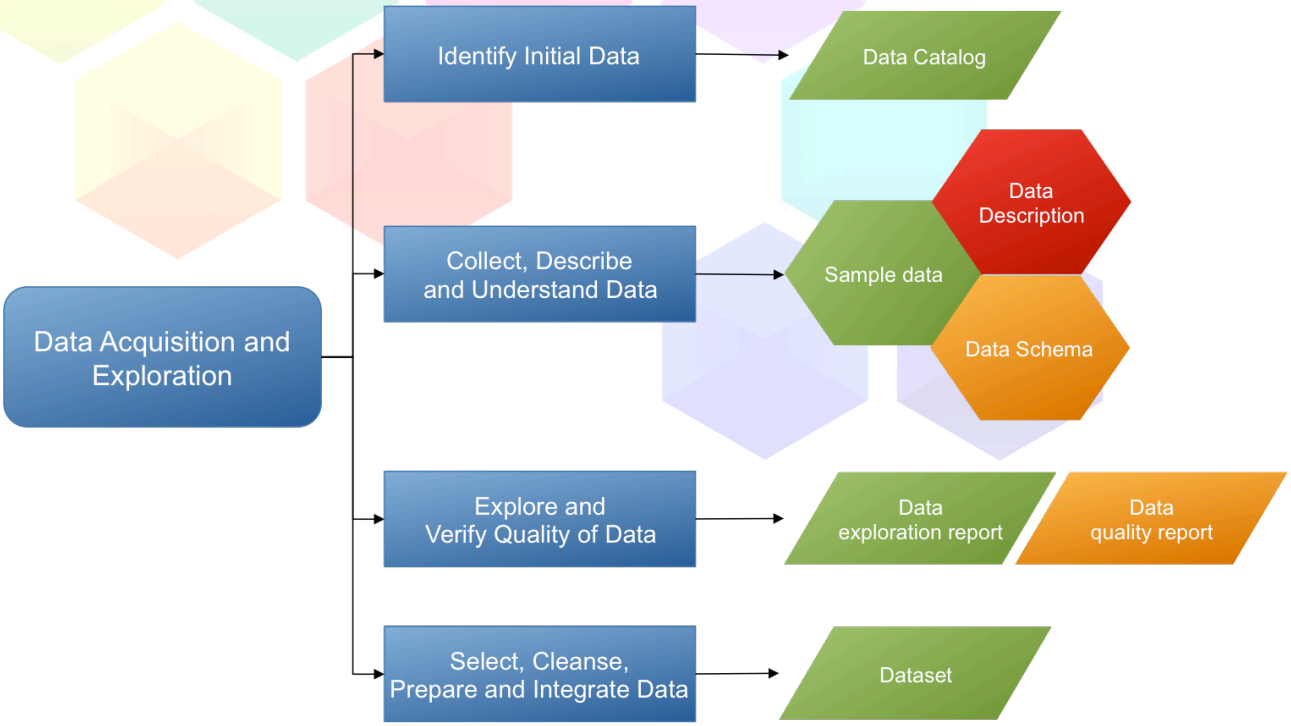

Illustration 7: Stages 3 – Business Understanding

This stage focuses on acquiring and exploring available data to gain better understanding on it, identifying data cleansing needs, identifying opportunities for data enrichment, and identifying analysis that can be done with the available data.

Activities and outputs of this stage are:

i. Identify Initial Data
   a. Data catalog

ii. Collect, Describe And Understand Data
   a. Sample data
   b. Data description

c. Data schema

iii. Explore and Verify Quality of Data

a. Data exploration report

b. Data quality report

iv. Select, Cleanse, Prepare and Integrate Data

a. Dataset

Stage 4: Analytical Model Development

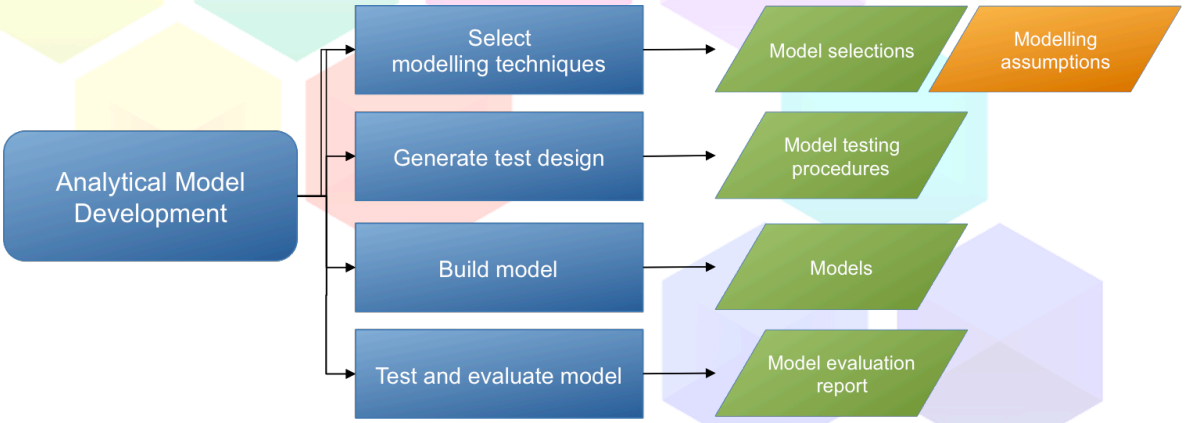

Illustration 8: Stages 4 – Analytical Model Development

This stage focuses on the development of data model and analysis algorithms to process data to produce results needed by the business.

Activities and outputs of this stage are:

i. Select modeling techniques
   a. Model selections
   b. Modeling assumptions

ii. Generate test design
   a. Model testing procedures

iii. Build model
   a. Models

iv. Test and evaluate model
   a. Model evaluation report

Depending on the results of the model testing and evaluation, development may go back to **Stage 3: Data Acquisition and Exploration** to acquire better datasets to improve the model

Stage 5: Data Product Development

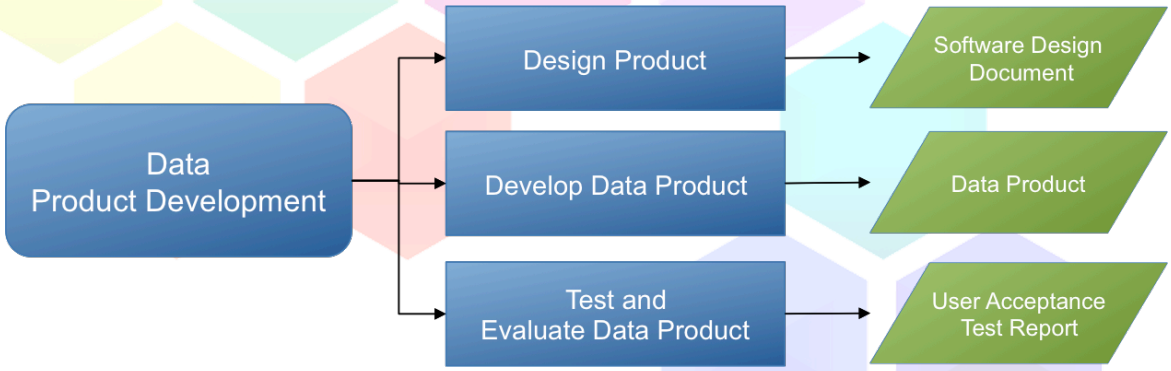

Illustration 9: Stages 5 – Data Product Development

This stage focuses in the development of Data Product, which can be a dashboard visualization or reporting software which displays the analysis result, or a more complex data driven application that utilizes the analyzed to handle specialized business needs.

Activities and outputs of this stage are:

i. Design product
    a. Software design document

ii. Develop Data Product
    a. Data product

iii. Test and Evaluate Data Product
    a. User acceptance test report

Stage 6: Transition to Production

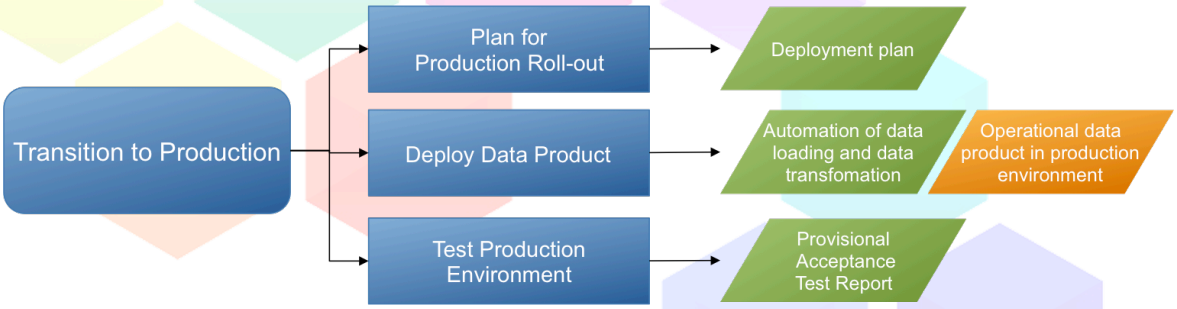

Illustration 10: Stages 6 – Transition to Production

After data product have been fully developed and tested, it is then evaluated against the business requirements, and then rolled out into the production environment with access to more data. Scheduling and automation of analysis processes from production data are also configured during this phase.

i. Plan for Production Roll-out
   a. Deployment plan

ii. Deploy data product
   a. Automation of data loading and data transformation
   b. Operational data product in production environment

iii. Test Production Environment
   a. Provisional Acceptance Test report

Stage 7: Monitoring

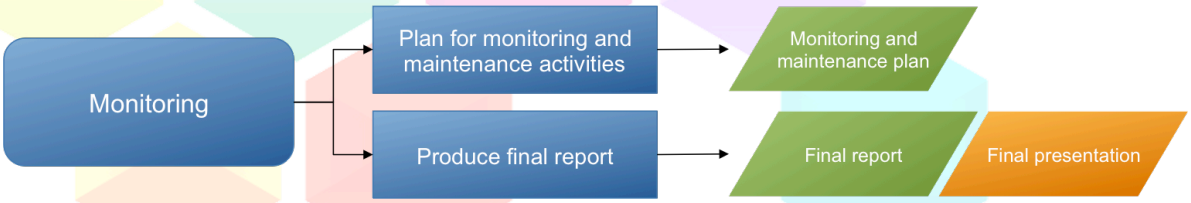

Illustration 11: Stages 7 – Monitoring

The project is then monitored for its effectiveness, stability and capacity with regards to business requirements. Any opportunities for further improvement and enhancements are recorded for planning of the next cycle of improvement.

Activities and outputs for this stage are:

i. Plan for monitoring and maintenance activities
   a. Monitoring and maintenance plan

ii. Produce final report
   a. Final report
   b. Final presentation