# 

# 

# An incubated co-working space for technology innovation

## Internal Document

**Binary Proposal Document**

**Table of Contents**

[1](#_Toc419677050)

[An incubated co-working space for technology innovation 1](#_Toc419677051)

[Internal Document 1](#_Toc419677052)

[Document Change Log ii](#_Toc419677053)

[Definition of Terms iii](#_Toc419677054)

[Description of Acronyms iv](#_Toc419677055)

[1 Document Overview 5](#_Toc419677056)

[1.1 Introduction 5](#_Toc419677057)

[1.2 Problem Statement 6](#_Toc419677058)

[2 Proposed Solutions 7](#_Toc419677059)

[3 Architecture 8](#_Toc419677065)

[4 Tools 9](#_Toc419677065)

# Document Change Log

This is a working document, which will be maintained with time. Team members, please ensure that any changes are recorded in the change log below – this is to ensure that each team member is always clear about which changes have been made and when.

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Description** |
| 0.01 | 21 March 2015 | Malusi Gcakasi | Created initial template document. |
|  |  |  |  |

# Definition of Terms

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Central Team | Team physically located at The Cortex Hub, East London |
| Remote Team | Team physically located outside of The Cortex Hub, East London |

# Description of Acronyms

|  |  |
| --- | --- |
| **Acronym** | **Description** |
| CH | The Cortex Hub |
| CT | Central Team |
| RT | Remote Team |

# Document Overview

## Introduction

# Big Data means is a collection of data sets, perpetually evolving large amounts of structured, semi-structured and unstructured data and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

# It is the result of practically everything in the world being monitored and measured, creating data faster than the available technologies can store, process or manage it.

# Big Data results in three basic challenges: *storing*, *processing* and *managing* it efficiently.

# *Storing:*

# A vast majority of Big Data is either duplicated data across various storage nodes or synthesized data, this serves as a precaution measure.

*Processing:*

Because of the large scale/amount of this data, retrieving it at the desired fast rate becomes the biggest challenge, this hinders the “processing” and analysis process that takes place before storage.

*Managing:*

As the scale of the data and systems progresses managing daily accumulated data is becoming difficult using basic database management tools.

* 1. **PROBLEM STATEMENT**

Public department “Home Affairs” has basic databases situated for/in individual towns and cities. As organised and minimal as this system may appear to be, it has a few flaws that prolong and hinder with a few processes.

These databases are not systematically synchronised in all these respective provinces, it becomes a major challenges to process, do mass analysis and reports since these databases exists independently.

Population accelerates/ascends on a daily to be precise which means regular updates are required for the system to be able to keep with accumulation of the population or the opposite. The Home Affairs currently faces obstacles in this regard.

A recent development in the USA by the state president has been made, where he ensured that Big Data for public entities is easily accessible to the citizens. The motive behind was to get ordinary citizens especially programmers to assess it, play around with it and explore it to detail so that solutions could be found/made in the process of manipulating the available data to them.

* 1. **SCOPE**

1. **PROPOSED SOLUTIONS**

To remedy the storage problem, parallel computing is the best solution. Different clusters and nodes can be created and located across the country whilst having one singe Master cluster connected to them all. This make it easy to modify and make additions via this Home Affairs cluster that would exist, in short all the processing and storage is done concurrently.

This data will still be in replicated and stored three times for precautions and all this computations will be done through cloud computation.

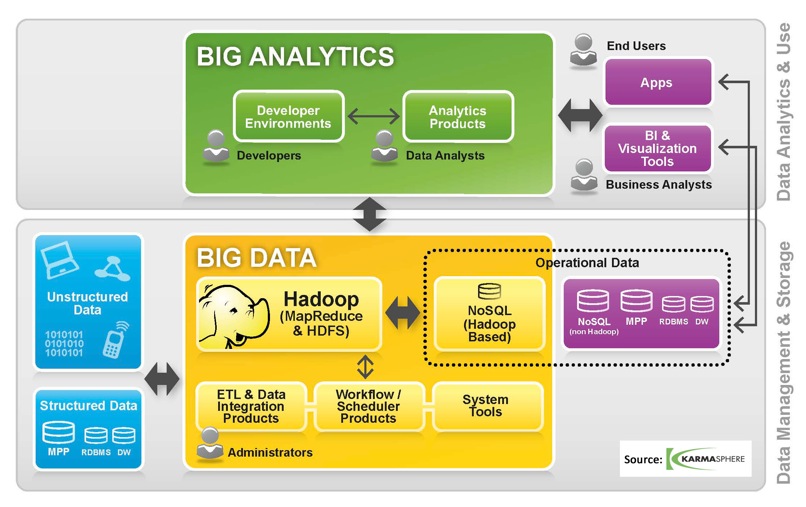
The existence of the cluster makes analysis and reporting easy as the large scale of this data can easily retrieved and accessed.

The extension of accessibility to an ordinary South African would allow a programmer to write an algorithm and system to handle calculations and regular updates of births and deaths in the country. This could be done accurately with useful information provided by the Department of Health.

This then diminishes the use of Census held every half a decade, and saving states resources and time.

Using Hadoop Cloudera’s ecosystem this is doable and possible, with the use of Mahout, Map reduce and other components.

1. **Architecture**



1. **Tools**

* Hadoop
  + *Data Storage*
    - Hadoop Distributed File System (HDFS)
  + *Data Processing*
    - MapReduce 2.0
  + *Data Access*
    - Mahout
    - Hive
  + *Data Management*
    - Oozie
    - Flume