

# Analysis of Dual Core Data

## Executive Summary

The objective of the project is to provide recommendations to improve the profitability of sales and delivery of Dual Core by harnessing the power of Big Data. In order to achieve this, we will be integrating the given data with two external data sources.

The Dual Core dataset has a relational data model. It consists of data related to customers, products, suppliers and order details. This data will be analyzed to understand the trends in sales and identify any potential issues that might be affecting sales.

In order to build our mediated schema, we will integrate the Dual Core dataset with two external data sources - the US Zip Codes database and the US holiday calendar database. Our mediated schema will create formatting rules for the data and rules for how we intend to identify duplicate rows and consolidate those rows into a single record. We will use HDFS as our database for records and Sqoop to import the data.

The Zip Codes database consists of location details like – City, State and Zip Code of all the locations in United States. By integrating this data with our DualCore, we are planning to identify the locations with least activity and therefore, come up with recommendations on where to set up delivery hubs to expand the business. By integrating the US holidays calendar database, the flow/trend of orders generated will be analyzed, based on which recommendations on discounts, order quantity, and product categories will be presented to improve the sales.