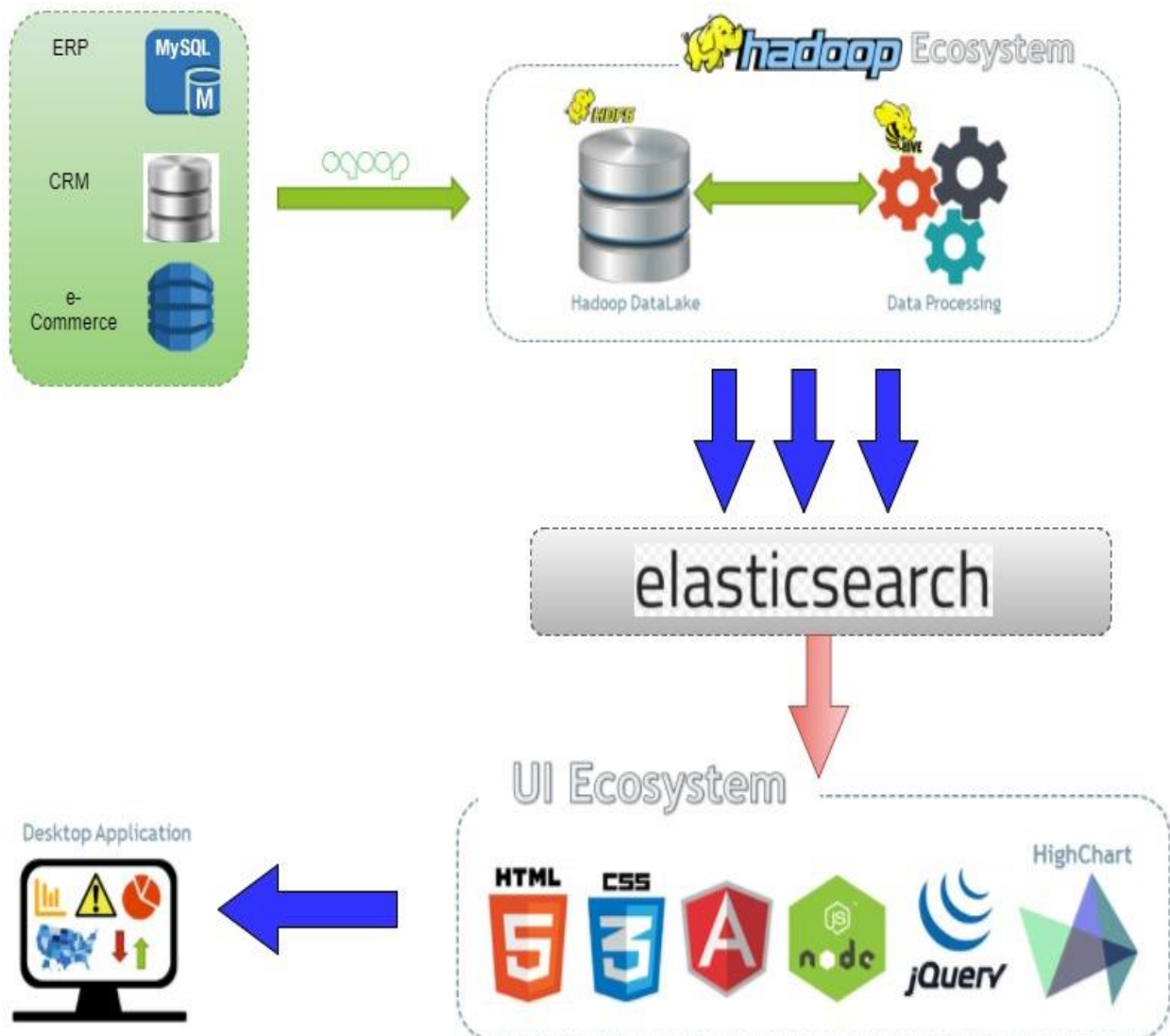# Big Data

High Performing Reporting Architecture on Big Data

Big Data - Is it really a buzz word Or is it worth it? Statistics have revealed that data driven decision for industry could potentially open avenues for huge growth. We are seeing a trend in the industry where traditional analytics of learning from sampling of data is being replaced by big data techniques as systems grow stronger. I am here to present a high performing architecture on based on big data.

In above architecture, consider we want to bring data from various front-end systems like ERP, CRM & e-Commerce into a common data platform where analytics could be perform to produce scalable and robust reporting solutions.

Here, I am considering all front-end systems storing data into RDBMS Databases like MySQL, Oracle, SQL Server etc.

Using Sqoop - an effective tool provided by Hadoop Ecosystem, we can write simple SQL queries to pull data into Hive. Hive is a data warehousing infrastructure based on Apache Hadoop. Hive is designed to enable easy data summarization, ad-hoc querying and analysis of large volumes of data. It provides SQL which enables users to do ad-hoc querying, summarization and data analysis easily.

Note: Hive is not designed for online transaction processing. It is best used for traditional data warehousing tasks.

Once we have data in hive, we can write SQL Like analytical queries on the data set and result set of those analytical outputs could be pushed to another datastore called Elasticsearch. Here, I am using Elasticsearch because it is a powerful document store which is extremely fast and provides a powerful search APIs. In order to push result set of Hive to Elastisearch, we can use a connector provided by elastisearch that does this job seamlessely.

After data is available in Elasticsearch, we can now use this data in front end using reporting javascript libraries like Highcharts to display the data in form we prefer.

Front-end stack that is represented in this architecture comprises of HTML, CSS, AngularJS as front-end scripting and Nodejs to be used in Backend APIs.