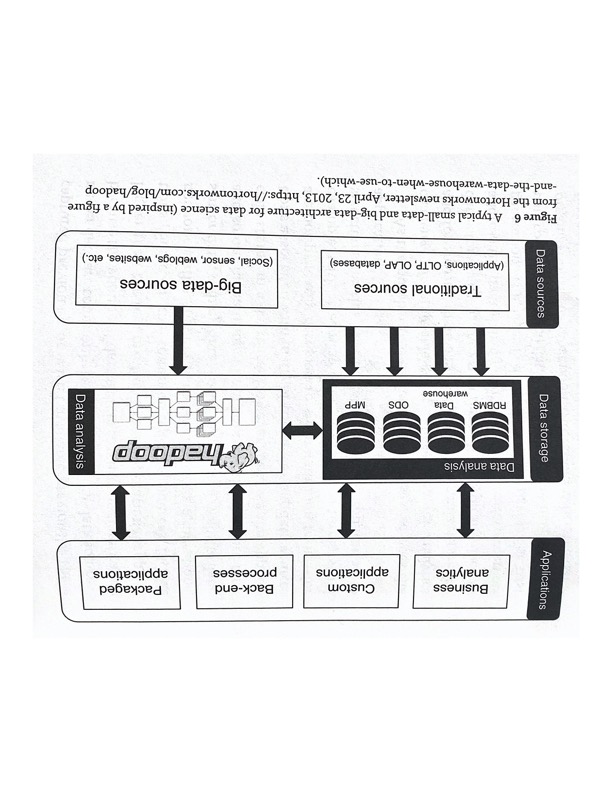
**Data science ecosystem**

It is very useful for mid and very large organizations.

****

Data science ecosystem helps generating valuable insights of data for any organization.

The data science ecosystem has four main components.

1. Data sources
2. Data storage
3. Data analysis
4. Applications

**Data sources-** Data sources has been generated from different sources like RDBMS- The traditional systems are often a backbone of the business intelligence (BI) solutions.  
Depending on maturity level of BI architecture, it can consist anything From basic copy of an operational application to an operational data store (ODS) to massively parallel processed PIPP BI database solutions & data warehouses.

**Data Storage-** Data washhouse is best understood process of data aggregation of analysis with goal of supporting decision making, most of data scientist time goes in finding, aggregating and cleaning the date prior to their analysis:  
It is possible to do data science without centralized of data repository it involves dumping the data from multiple operational databases into a single database. ETL- extraction , transformation and load ETL processes. The tools used to support mapping, merging & movements of data between databases online analytical processing (OLAP) used to describe these operations

OLAP systems are useful for data exploration and reporting they don’t enable data modeling for automation extraction of Pattern from Data. once the data from across the organization has been aggravated and analyzed within the bee BI systems this, this analysis Can then be used as a input to range of customers an application layer.

**Data Analysis-**

Hadoop is core component for this ecosystem and best known as bigdata processing framework. This framework was developed by Apache Software foundation.

How the Hadoop does the processing of data? MapReduce is core of Hadoop and it implements split-apply-combine strategy

1. A large data sets it split up into separate chunks and data it stored on a different nodes in the cluster
2. MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).
3. MapReduce runs on each node and produce the final result after combining.

Over the past couple of years the Hadoop platform is being used as an extension of enterprise data warehouse. Most recent data store in data warehouse & old data in a data Hadoop dusters.  
Most of the enterprise-level database, and it has features that connect the data warehouse & Hadoop allowing data scientist, using SQL to query the data in both places as if they all are located in one environment.

**Applications-** As part of applications, Organization runs the different business applications, Custom applications, Backend process, and packaged applications.

**Big Data Infrastructure and there programing languages**

Hadoop is best known as big-data processing frameworks. But it is mostly used for batch processing of data. Batch processing is appropriate where dataset is static during the processing and where the results of the processing are not required immediately.

Other big-data processing include Storm, Spark, and Flink. In steam processing each element in the stream is processed as it enters the system and consequently the processing operations are defined to work on each individual element in the steam rather than on the entire data set.

Spark and Flink are both Hybrid processing good frameworks. Spark is a fundamentally a batch processing framework, similar to Hadoop, But also has some stream oricessing capabilities whereas Flink is a stream processing framework that can also be used for batch processing.