# KEY TECHNOLOGIES BEHIND BIG DATA



Big data has been making a big splash for almost a decade now, but most people are still uncertain about what “big data” actually is, how it works, and the technologies that lay behind its magic. To

answer these questions, I am publishing an extensive, 6-part blog series on big data technologies and what it means to you.

To illustrate the power of big data in business, let's look at a popular example from the retail giant Walmart. According to WalmartLabs' Stephen O'Sullivan, back in 2013 Walmart made a pivotal move by collating all of its data resources onto a single, 250-node Hadoop cluster that acts as a big data mega store. Utilizing several types of traditional and big data technologies, this centralized big data hub processes *new data* on a scale of terabytes per day, as well as *historical data* on a scale of petabytes per day. Similarly, as part of WalmartLabs' efforts to advance itself via big data analytics, it developed a new semantic search engine called Polaris. With respect to the search engine's effectiveness, Walmart's Corporate website reported that "Walmart.com has already seen an approximate 10-15 percent increase in shoppers completing a purchase after searching for a product using the new search engine".

While this may seem like something old, we continue to learn about Big Data through successful trial and error followed by production implementation. **This first post in the series will cover how "big data" is defined and some of the technologies that are commonly used for handling it.**

# The Simple Definition of Big Data

Any introduction to big data would be incomplete without discussing the most common 3-Vs talked about with Big Data. Big data is data that has volume, variety or velocity such that it can't be handled and processed using traditional data technologies. This is called the "3-V criteria". If your data meets any of the 3-V criteria, then it meets the big data threshold and would be reasonable to classify your project as a "big data project".

The *3-V criteria* is as follows:

- **Volume**: If your organization owns at least 1 terabyte of data – about the storage capacity it would take to save 500, 2-hour movies – then big data technologies are likely to be in order.
- **Velocity**: Whether your organization is processing data in batch, real-time or near-real time, if data is entering your IT system at a rate of 30 Kbps to approximately 30 Gbps, then you should be using big data technologies to meet these requirements.
- **Variety**: If your organization uses multi-structured data, then big data technologies are required for efficient data storage and handling. *Multi-structured data* is any combination of data that comes in structured, semi-structured or unstructured formats.

# Some Key Big Data Technologies

Now that you understand the circumstances from which big data projects arise, it's a good time to get familiar with some of the common types of technologies that are used in big data solutions. Some common technologies in the big data ecosystem are:

- **Apache Hadoop**
- **Apache Hive / Apache Pig**
- **Apache Sqoop**
- **In-memory Databases**
- **NoSQL Databases**

- **MPP Platforms**

**The "Hadoop" Ecosystem**
In general, when a person uses the term "Hadoop" to describe their solution, they are referring to the Hadoop ecosystem, which is commonly comprised of Hadoop HDFS, YARN, and MapReduce. HDFS is the *Hadoop Distributed File System* that's used for storing large volumes of multi-variety big data onto clusters of commodity servers. MapReduce is responsible for the parallel processing of data that sits on the HDFS and YARN (*Yet Another Resource Negotiator*) handles Hadoop's resource management requirements.

**Apache Hive and Apache Pig**
Native MapReduce jobs are written in Java. Hive is a SQL-like querying language that allows users to query data from Hadoop while by-passing Java programming requirements. Pig is a procedural language that simplifies the process of writing MapReduce jobs and allows users to both query and explore the data that sits in Hadoop. Hive and Pig share common goals. These are:

1. To bypass Java programming requirements when writing MapReduce jobs, and
2. To open up big data resources to a broader spectrum of professionals

**Apache Sqoop**
Sqoop is a tool for transferring data between Hadoop and traditional database solutions.

**In-memory Databases**
In-memory appliances, like Apache Spark, are used for generating big data analytics from real-time or near real-time streaming data.

**NoSQL Databases**
NoSQL databases are databases specifically designed for storing and processing the multi-variety data that characterizes big data. There are four main types of NoSQL database. These include:

1. Column stores
2. Document databases
3. Key-value stores, and
4. Graph databases

**Massively Parallel Processing (MPP) Platforms**
MPP platforms are a parallel-processing alternative to MapReduce. MPP platforms run off of custom hardware, making them the more costly alternative.