# 01. Data Engineering Described

Data Engineer's goals:

- `produce optimum ROI` and reduce costs (financial and opportunity)
- reduce risk (security, data quality)
- maximize data value and utility
- must constantly optimize along the axes of cost, agility, `scalability`, simplicity, reuse, and `interoperability`.

## History of data engineering

- Bill Inmon invented the term `data warehouse` in 1989.
- IBM developed the relational database and `Structured Query Language (SQL)`.
- Oracle popularized this technology.
- `Massively parallel processing (MPP)` is a first age of scalable analytics, which uses multiple processors to crunch large amounts of data. Relational databases were still most popular.
- Internet companies like Yahoo or Amazon: after internet boom all of those companies looking for new systems, that are cost-effective, scalable, available, and reliable.
- Google published a paper on the Google File System and `MapReduce` in 2004. It starts ultra-scalable data-processing paradigm.
- Yahoo: based on Googles work, they develop Apache Hadoop in 2006.
- Amazon created `Amazon Web Services (AWS)`, becoming the first popular public cloud.
- Hadoop based tools like Apache Pig, Apache Hive, Dremel, Apache HBase, Apache Storm, Apache Cassandra, Apache Spark, Presto and others are becoming very popular. Traditional enterprise-oriented and GUI-based data tools suddenly felt outmoded.
- `Hadoop ecosystem` including Hadoop, `YARN`, `Hadoop Distributed File System (HDFS)` is a king in late 2000s and in the beggining of 2010s.
- `Apache Spark` rise because too many tools on the market drove to inventing one unified tool, which was Apache Spark. It got very popular in 2015 and later.
- Simplification. despite the power and sophistication of open source big data tools, managing them was a lot of work and required constant attention. data engineers historically tended to the low-level details of monolithic frameworks such as Hadoop, Spark, or Informatica, the trend is moving toward `decentralized, modularized, managed, and highly abstracted tools`.

## Data team

Upstream stakeholders:

- Data architects

- Software engineers
- `DevOps engineers`

Downstream stakeholders:

- Data scientists
- Data analysts
- Machine learning engineers and AI researchers

## Data maturity

Data maturity is the progression toward higher data utilization, capabilities, and integration across the organization.

Three stages:

- starting with data
- scaling with data
- leading with data