**Data profiling**:Data profiling is the process of examining the data available in an existing data source (e.g. a [database](http://en.wikipedia.org/wiki/Database) or a [file](http://en.wikipedia.org/wiki/Computer_file)) and collecting [statistics](http://en.wikipedia.org/wiki/Descriptive_statistics) and information about that data. The purpose of these statistics may be to:

* Find out whether existing data can easily be used for other purposes
* Improve the ability to search the data by [tagging](http://en.wikipedia.org/wiki/Tag_(metadata)) it with [keywords](http://en.wikipedia.org/wiki/Keywords), descriptions, or assigning it to a category
* Give [metrics](http://en.wikipedia.org/wiki/Software_metric) on [data quality](http://en.wikipedia.org/wiki/Data_quality), including whether the data conforms to particular standards or patterns
* Assess the risk involved in [integrating data](http://en.wikipedia.org/wiki/Data_integration) for new [applications](http://en.wikipedia.org/wiki/Data_profiling), including the challenges of [joins](http://en.wikipedia.org/wiki/Join)
* Assess whether [metadata](http://en.wikipedia.org/wiki/Metadata) accurately describes the actual values in the source database
* Understanding data challenges early in any data intensive project, so that late project surprises are avoided. Finding data problems late in the project can lead to delays and cost overruns.
* Have an enterprise view of all data, for uses such as [master data management](http://en.wikipedia.org/wiki/Master_data_management) where key data is needed, or [data governance](http://en.wikipedia.org/wiki/Data_governance) for improving [data quality](http://en.wikipedia.org/wiki/Data_profiling).

Data profiling is an analysis of the candidate data sources for a data warehouse to clarify the structure, content, relationships and derivation rules of the data. Profiling helps to understand anomalies and to assess data quality, but also to discover, [register](http://en.wikipedia.org/wiki/Data_profiling), and assess enterprise metadata. Thus the purpose of data profiling is both to [validate](http://en.wikipedia.org/wiki/Data_profiling) metadata when it is available and to discover metadata when it is not. Data profiling is one of the most effective technologies for improving data accuracy in corporate databases.

**Reasons of Data Profiling**:

1) To check that the data is suitable for other business requirements then one from where the data originated.

2) In case of migration from one system to another, it is an exercise to analyse the data if it is suitable for the target.

3) Profiling can result the information about its container, whether the database or file system is suitable enough to hold the information or it needs change.

4) To categorise the data to use it more effectively to make effective business decision.

5) Sometime profiling is required to check the quality of the data. Quality check of the data can yield wonderful results and could be beneficial for DSS and other interfaces.

**Data profiling tools:**

**1.)Trillium Discovery (TS discovery):** TS Discovery is the automated data profiling and data discovery component of the Trillium Software System, a robust, scalable, highly available and easily deployable solution for mission-critical enterprise data quality. TS Discovery enables business and data analysts to assess large volumes of data within and across systems collaboratively. Robust data profiling capabilities allow users to:

* Understand data [domains](http://www.trilliumsoftware.com/home/products/TSDiscovery.aspx) and formats within the data in preparation for [data migration](http://www.trilliumsoftware.com/home/products/TSDiscovery.aspx) and data integration initiatives
* Ensure that data will support business requirements
* [Validate](http://www.trilliumsoftware.com/home/products/TSDiscovery.aspx) data conformity to specific business rules and defined data standards.

TS Discovery routinely assesses data to ensure that high quality is maintained at all times and monitors production systems for anomalies.  Delivering a more comprehensive understanding of enterprise across systems translates to more accurate project planning, scoping, and feasibility assessment and reduces risk in meeting daunting project deadlines and aggressive budgets.

The Trillium Software System is a fully integrated, total data quality suite that delivers a single user experience for complete, global data quality life cycle management across the enterprise.  Purpose-built for seamless movement to and from views related to each phase of the [data quality management](http://www.trilliumsoftware.com/home/products/TSDiscovery.aspx) life cycle, the Trillium Software System emphasizes up front and ongoing investigation and improvement (process design, development and test phases).

**2.)Informatica Data Quality:** Informatica Data Quality ensures that authoritative and trustworthy data is available to all stakeholders and data domains, for all projects and business applications—on premise or in the cloud—using a single, unified platform. It transforms the way your business works by instilling trust in all your data, for all your needs, at all times.

**3.)Data stage:** WebSphere DataStage performs data transformation and movement from source systems to target systems in batch and in real time.

**4.)IBM WebSphere Information Analyzer** is a data profiling and analysis tool that is a critical component of IBM Information Server. It is designed to help business and data analysts understand the content, quality, and structure of their data sources by automating the data discovery process.