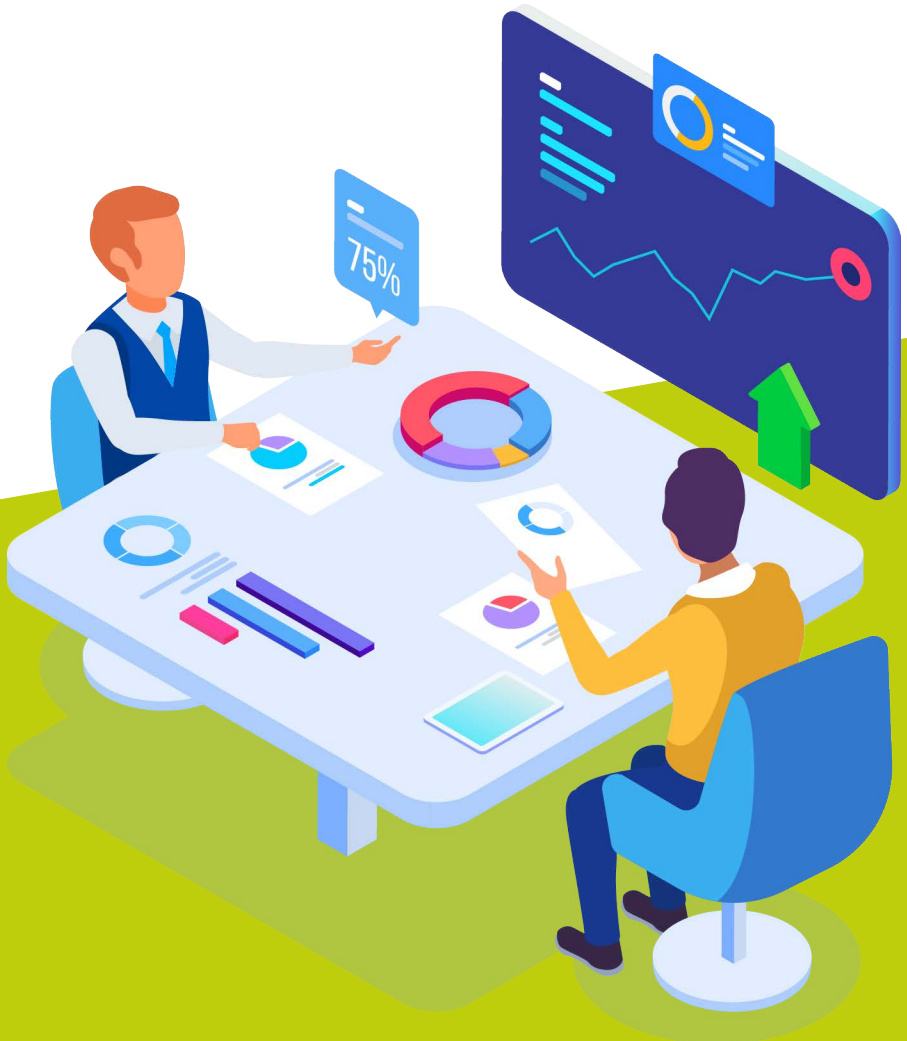
HOW TO BUILD A BUSINESS INTELLIGENCE SYSTEM



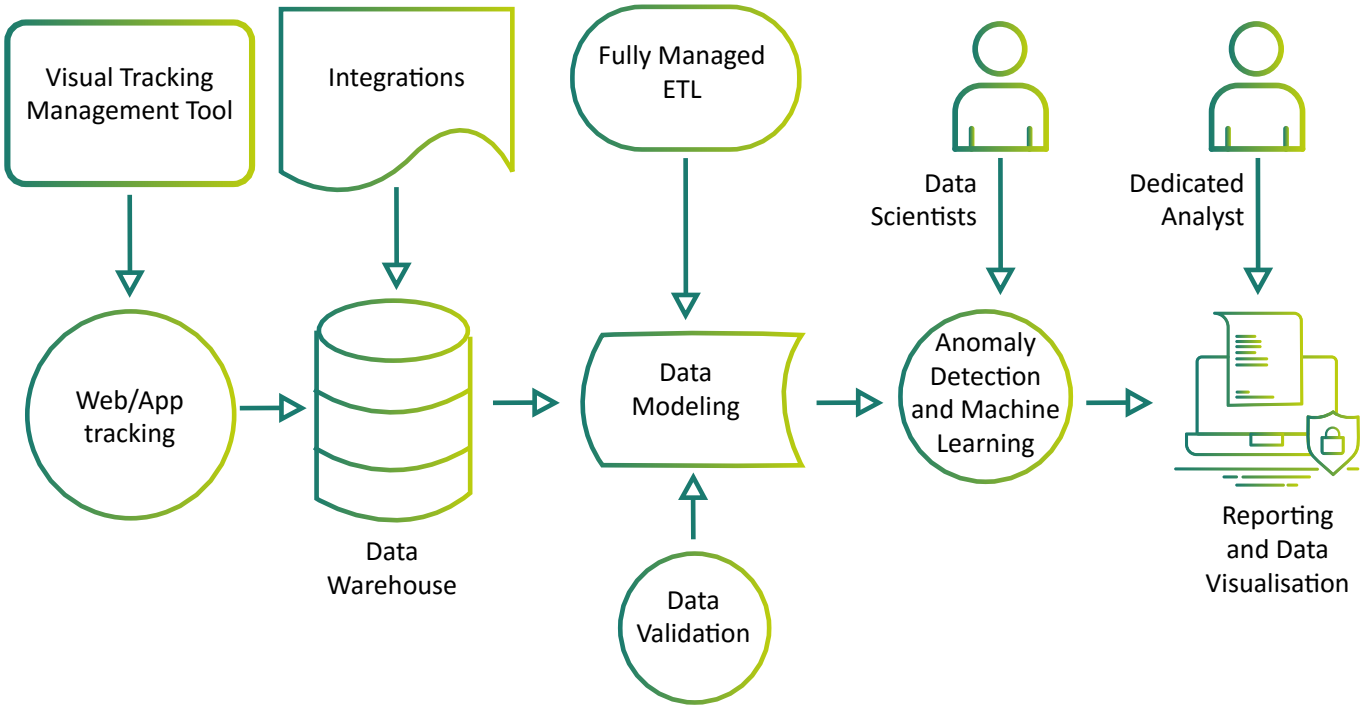
How to Build a Business Intelligence System

Business intelligence systems come in all shapes and sizes, but, at their core, they all have the same basic components which perform distinct functions within the system as a whole.

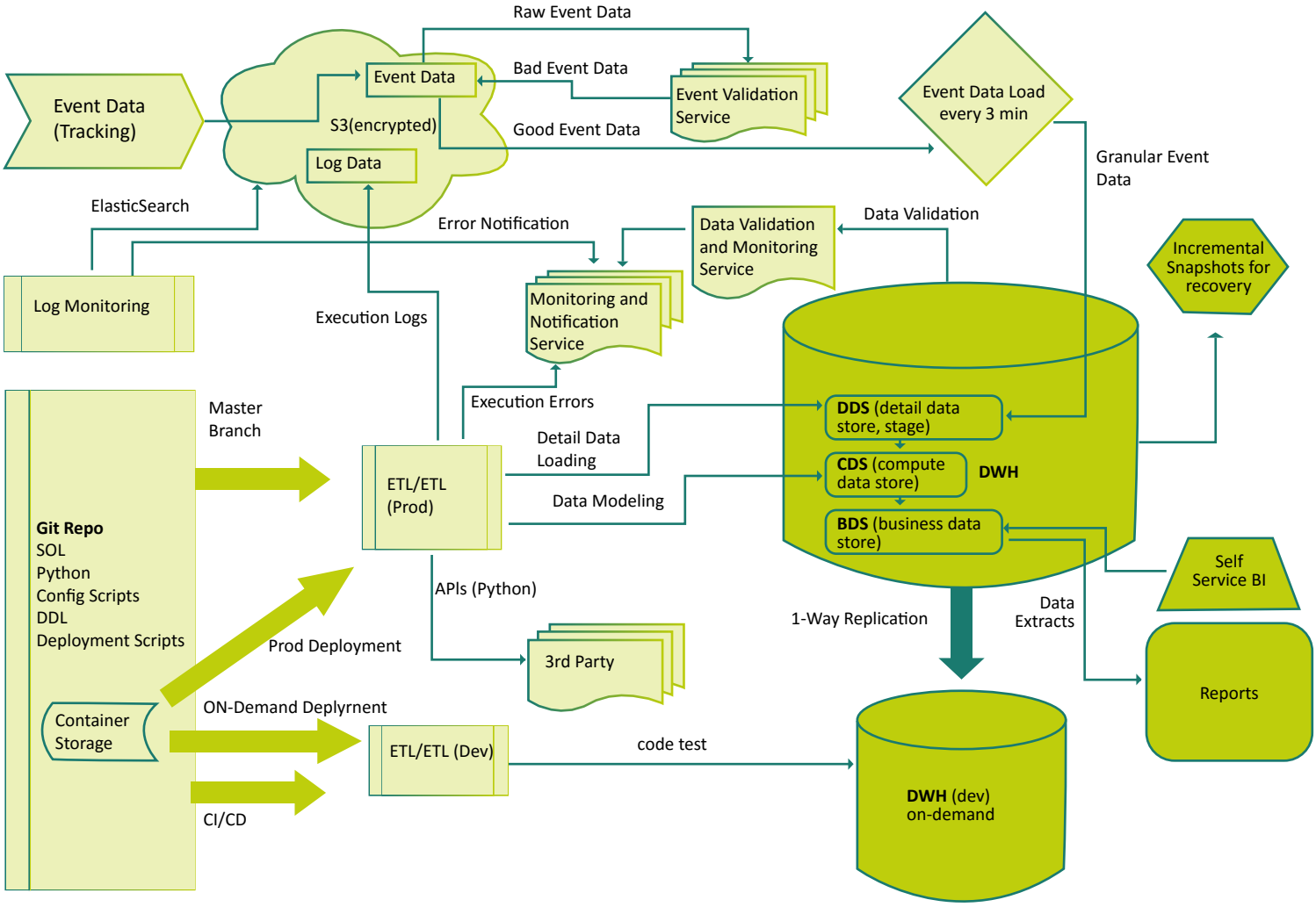
These components are:

- Data integration
- ETL/ELT
- Tracking
- Data warehouse (with or without a data lake)
- Data modeling (machine learning, data science, forecasting)
- Reporting and data visualizations

A basic business intelligence system can look something like the diagram below:



However, it can have a more sophisticated design, with data validation, disaster recovery, as well as logging added into the BI system. The chart below illustrates a more desirable design that will scale with the business without the need to change the major components as the company grows.



Data Integration

Data integration refers to the need for a business to connect to or extract data from the outside sources. This is usually done with the advertising data from the platforms like Google and Facebook. However, as the business grows, it strives to integrate all the data that may have analytical value, even HR data from HR platforms used by the business. The process of integration refers to the need of the business to extract data from external systems and bring that data into the business intelligence system. This is often done incrementally, where the integration process will extract the data from the external provider, usually data covering the last 24 hours, bring that data into the data warehouse, and merge that data into historical data from the same provider already stored in the data warehouse.

ETL/ELT

The ETL/ELT process defines all processes, including the API calls that are made to external data stores, like Facebook and Google.

- ETL stands for extract, transform, load.
- ELT stands for extract, load, transform.

ETL is part of a more classical BI system, which has its beginnings in the early 2000s, where transformation of the data was done while the data was being moved between many systems. Over time, as the data sizes grew, transforming the data on the fly meant that mechanisms that were responsible for data transfer needed to have stronger computational resources to process large amounts of data. This was not scalable, and many companies switched to the ELT approach, where the transformation of the data was done inside the data warehouse or data lake after the data was loaded, because both of these systems were designed for manipulating large amounts of data.

There is a proliferation of ETL/ELT tools available. Some of the more popular are:

- Talend
- Pentaho
- Airflow
- Matillion
- AWS Glue (Amazon Web Services)

Tracking

Gathering of data, often referred to as “tracking” has become a key element of a healthy BI system because it enables a company to better understand how visitors and registered users interact with the company’s products or services on their website or inside their app. Many businesses try to save money by implementing rudimentary tracking themselves. This strategy always backfires as the business scales above 50 employees and the data needs grow. Tracking, done well, is such a complicated and resource-consuming process that it is best to outsource it to an external provider, who has a team of experts dedicated to doing tracking well. There are many tracking technology companies that provide a multitude of options in all price ranges, such as:

- Snowplow
- Google Analytics
- Amplitude (Product Analytics Tool)
- Kissmetrics
- Mixpanel
- Heap

One key feature that the business should always consider, when choosing a tracking solution, is whether it would like to have access to the detailed event data. This important difference will depend on whether the organization chooses to prioritize the immediate, self-serve product use cases, or is more concerned with building a robust, long-term data capability, of which product analytics is one (of many) use cases. If the latter is the case, then the organization should insist on having access and keeping their detailed tracking data as a strategic capability.

Exercise:

Calculate the cost of building an in-house tracking solution, not including the time and opportunity cost. For a bare minimum solution (i.e., SQL knowledge required and analysts available to use the system!) this is an estimate on costs to consider:

- *Senior engineers ($100-$200k annual salary, depending on region)*
- *Project manager ($60-80k annual salary)*
- *Platform cost estimate for 100M events per month:*
  - *Data visualization tool (e.g., Periscope) - +$100k annually*
  - *Data collector (e.g., mParticle) +$100k annually*
  - *SQL data storage (e.g., Amazon Redshift) - +$20k annually*
  - *Minimum cost (monetary only): $500.000*

Data Warehouse
(with or without Data Lake)

A data warehouse is a repository for structured data that is used by the business and data analysts to store and analyze data. It is considered to be a core component of the business intelligence system and serves as a central repository of integrated data from one or more disparate sources.

A data lake is a centralized repository that enables the organization to store all of its structured and unstructured data at any scale. The data stored in the data lake is usually as-is data, without having any unifying structure to it. Data lakes enable companies to run different types of analytics on the as-is data – from dashboards and visualizations to big-data processing, real-time analytics, and machine learning to guide better decisions. However, it is not as smooth to analyze the data across disparate sources in a data lake as it is in a data warehouse. This is why most organizations decide to have both, a data lake and a data warehouse.

Data Modeling
(Machine Learning, Data Science, Forecasting)

Data modeling refers to activities and processes that model and combine data. These may be, but are not limited to:

- Cleaning and sanitizing the data
- De-duplication of data
- Aggregating of data and rolling it up into time periods, such as years
- De-normalising the data, where dimensional elements appear in the fact tables In addition to the above standard data modeling processes, a company, depending on their level of data maturity, can choose to add these additional data modeling processes:
- Run machine learning algorithms
- Conduct data science experiments
- Forecast time-series data
- Classify and segment users

Regardless of the combination of activities, it is always important to ensure that higher-order functions, such as data science, are done after the lower-order functions, such as data cleaning, complete successfully. This ensures accurate results and consistency in the results of the higher order functions. Data modeling processes are usually applied to the data by data

analysts or data scientists. They are often automated to be executed daily overnight so that the results are ready for consumption by the business decision makers in the morning.

Reporting and Data Visualizations

Data visualization is the representation of data in the graphical form. It aims to simplify data for users to understand the meaning easily. From a technical perspective, it is important to choose the right tool, type of chart, style, and layout, which are all important elements of modern data visualization. What is even more important is to understand

- the purpose of a visual or a dashboard (the combination of related visual elements)
- what requirements business users have
- what questions they expect to find answers for
- what level of detail should be and so on – all such questions need to be discussed in order to create meaningful and insightful visuals rather than a good-looking, but useless charts.

Data visualization is a form of communication that simply portrays raw data in many different forms to make it easier for the user to extract information. It does so by giving it a graphical shape that makes it much neater and has an appeal to it. In order to create a good visualization, the main principles of data visualization need to be followed.

Some of these principles are:

- Accuracy - The visual must represent correct and full information. Misleading charts are not generally acceptable.
- Convenience - A user-friendly form of a visual report (styles, layout, readability etc.) must be achieved.
- Scalability - The ability to accommodate growing volumes of data and data sources without problem. This is also related to the maintainability of reports.
- Simplicity - Visualizations need to include only those elements that are really required for the report and bring value to it.
- Standardization - Often, visual reports are not a single chart, but a set of visuals or dashboards. They can be a part of boards, workbooks, stories etc. All these mentioned visualizations should have a standard form. This refers to the design, data structures, layouts, and so on.
- Interactivity - In most cases, interactivity is a good choice and allows achieving more insights from the visual and looking at the visual from different perspectives. Most of these principles must not be compromised, but some can be omitted for specific use cases. For example, if visualization will be used only as a static view (maybe as part of PowerPoint presentation) we can skip the “interactivity” principle.