**Data Workflow:**

**There are four general steps through which data flows within an organization:**

1. **Data collection & storage**: Collect and Ingest data, from web traffic, surveys, or media consumption for example.
2. **Preparation**: prepare it, which includes "cleaning data", for instance finding missing or duplicate values, and converting data into a more organized format.
3. **Exploration & Visualization**: explore, visualize, and build dashboards to track changes or compare two data sets.
4. **Experimentation & Prediction**: ready to run experiments, like evaluate which article title gets the most hits, or to build predictive models, for example, to forecast stock prices.

**Data Engineers:**

Data engineers are responsible for the first step of the process: ingesting collected data and storing it.

They are responsible for laying the groundwork for data analysts, data scientists and machine learning engineers.

If the data is scattered around, corrupted, and difficult to access, there's not much to prepare, explore, or experiment with.

And that's exactly why you need a Data engineer: their job is to **deliver the correct data, in the right form, to the right people, as efficiently as possible**.

They:

* Ingest data from different sources.
* Optimize the databases for analysis.
* Manage data corruption.

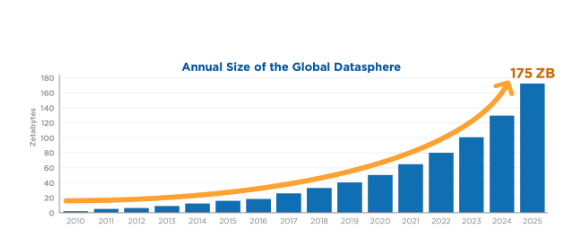
Data engineers develop, construct, test, and maintain architectures such as databases and large-scale processing systems to process and handle massive amounts of data.

**Big Data & Data Engineers:**

**With the advent of big data, the demand for data engineers has increased.**

Big data can be defined as data so large you have to think about how to deal with its size, because it's difficult to process using traditional data management methods.

This graph helps make sense of the growth of big data.

****

**The Five Vs:**

**Big data is commonly characterized by five Vs:**

1. **Volume**: the quantity of data points.
2. **Variety**: type and nature of the data (text, image, video, audio).
3. **Velocity**: how fast the data is generated and processed.
4. **Veracity**: how trustworthy the sources are.
5. **Value**: how actionable the data is.

Data engineers need to take all of this into consideration.

---------------------------------------------------------------------------------------------------------------------------------------------->>>