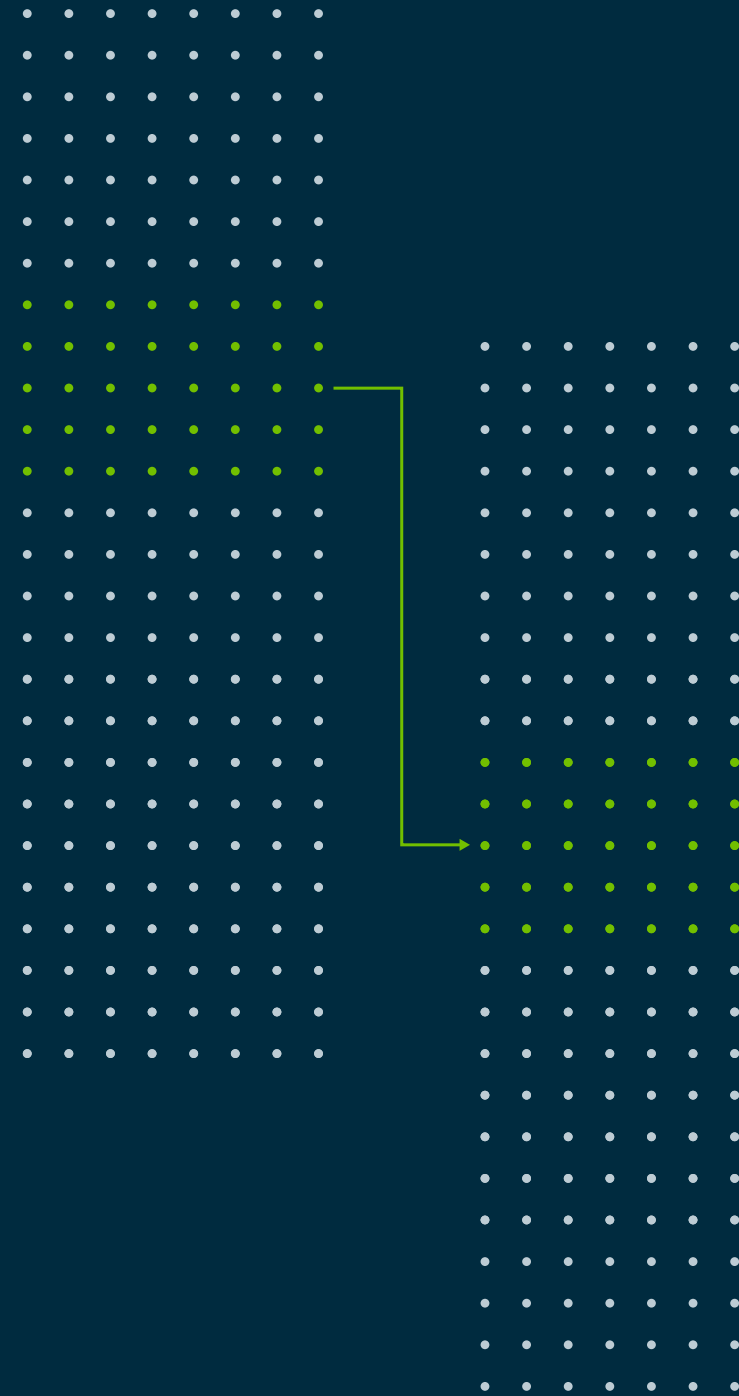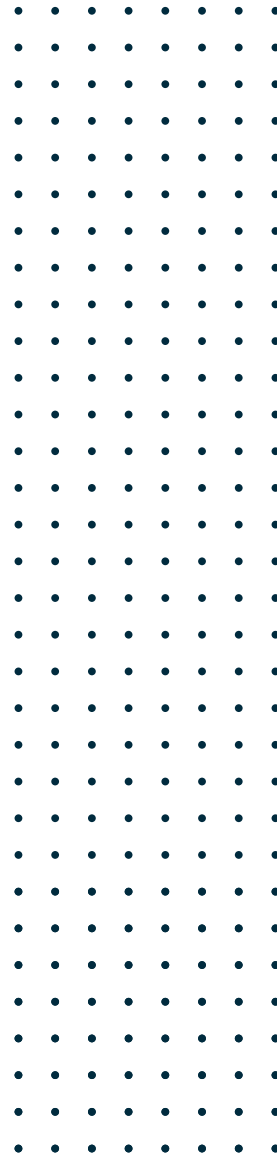Collibra

The complete guide to

Data Lineage

Data is everywhere.

Today, our every activity – from watching TV to going on a run – has the potential to leave a digital trace. In fact, every day, we generate over **2.5 quintillion bytes of data:**

- Google processes more than 40,000 searches every second

- We send 16 million text messages every minute

- Five new Facebook profiles are created every second

While these statistics are mind-blowing, what's even more staggering is that…

Source

90%

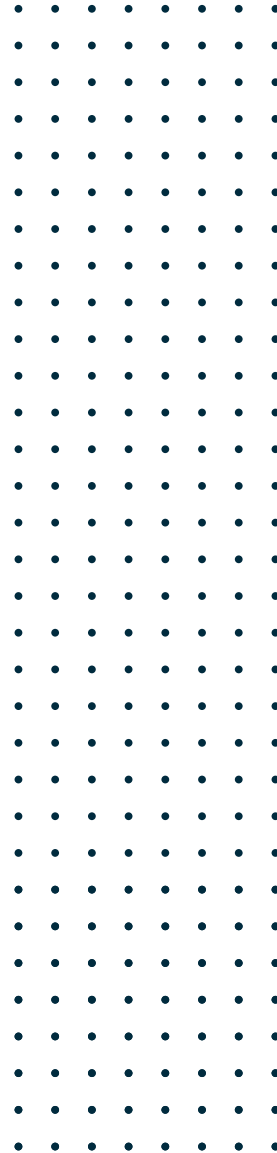of the world's data was produced in the last two years.
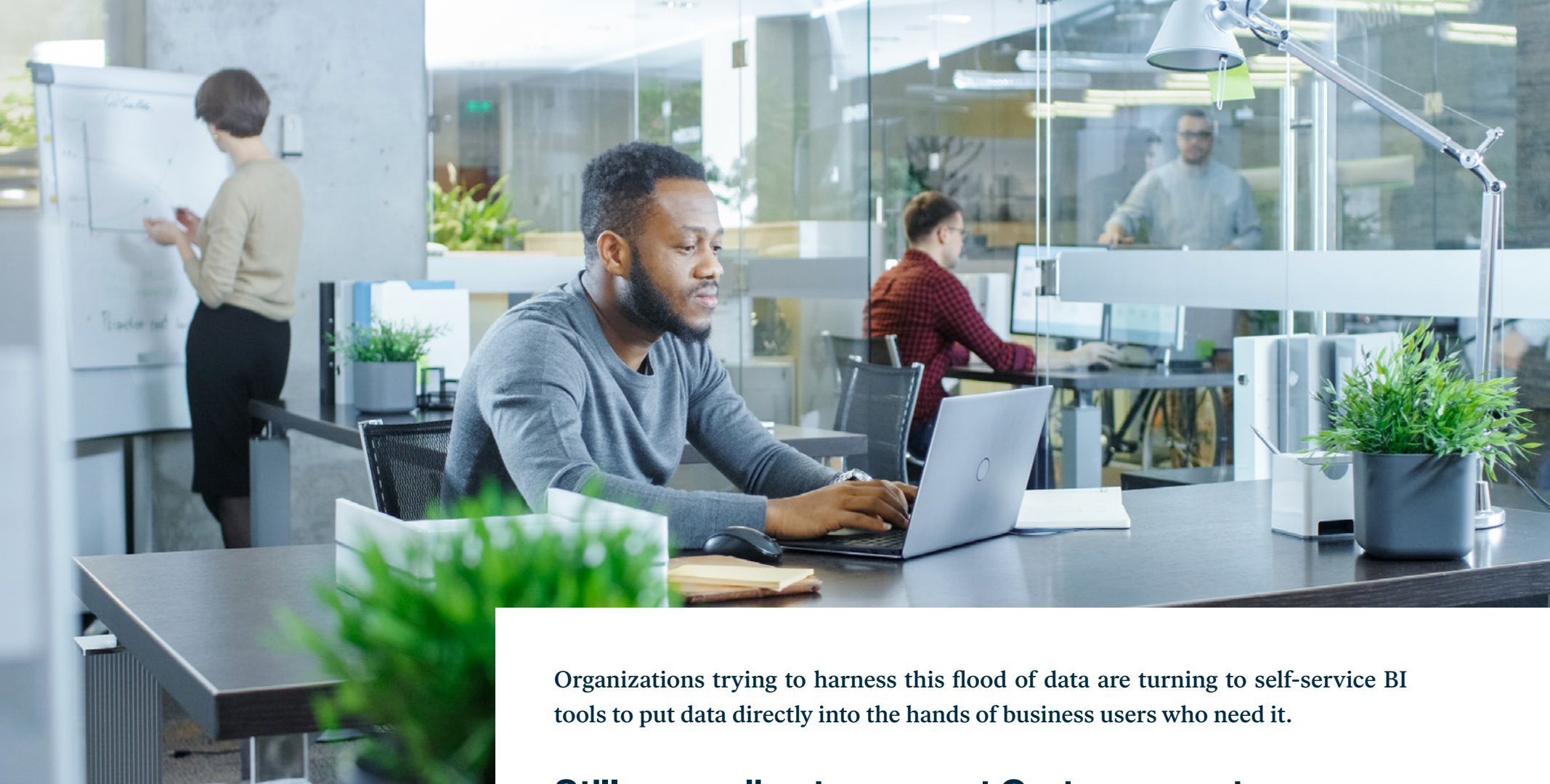
Source

While data is rapidly expanding, our capacity to derive value from the data is not keeping pace.

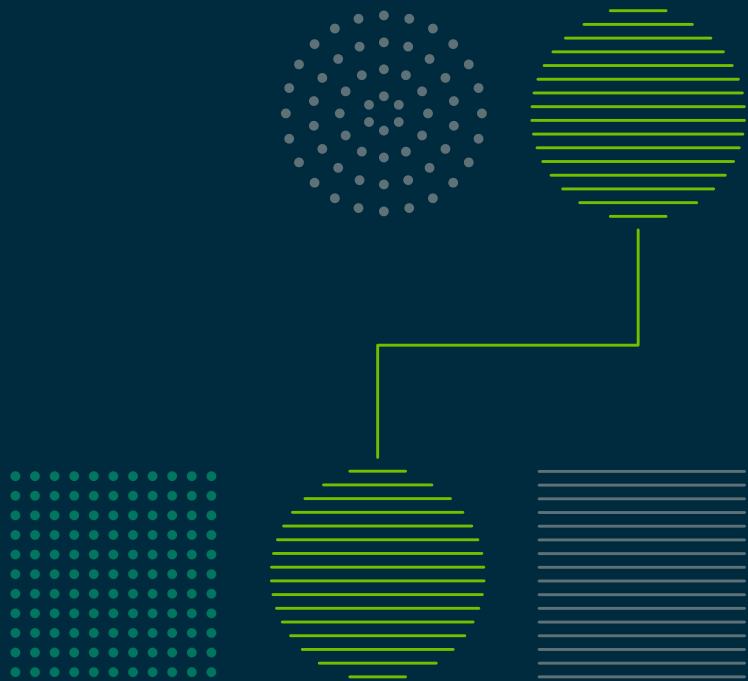**Most companies are capturing only a fraction of their data's value**

— Mckinsey

Organizations trying to harness this flood of data are turning to self-service BI tools to put data directly into the hands of business users who need it.

Still, according to a recent Gartner report more than 60% of those initiatives will fail to deliver trustworthy data.

To combat the data deluge and actually extract value from your data you must trust your data.

The solution?
Data lineage!

What is data lineage?

Data lineage describes how data transforms and flows as it's transported from source to destination across its entire lifecycle. It is a visual representation of the 5W's of data:

- Who is using the data?

- What information does it contain?

- When was the data created/transformed?

- Where does the data come from?

- Why does the data exist?

Data pulled from various sources sometimes ends up in a staging area – a data warehouse or a data lake. Before the data is delivered to the business user, it is typically aggregated, manipulated, and transformed through ETL tools, spreadsheets, ad hoc SQL or Python scripts. This manipulation of data can lead to inaccuracies in data and an overall lack of trust.

Luckily, data lineage restores this trust by providing lineage graphs that map out the flow of data as it moves from source to destination.

Data lineage graphs provide a roadmap of:

- **Data consistency**

- **Completeness**

- **Trustworthiness**

However, with data and data sources growing exponentially, mapping the movement of data across your enterprise can be extremely difficult and time consuming. This is why it is important to invest in an **automated lineage solution,** which automatically extracts technical lineage from source systems and maintains lineage diagrams.

What are the advantages of automated data lineage?

- Dramatically speeds up discovery to insight

- Enables IT to spend less time documenting data flows and more time on strategic initiatives

Automated data lineage benefits both IT and the business user:

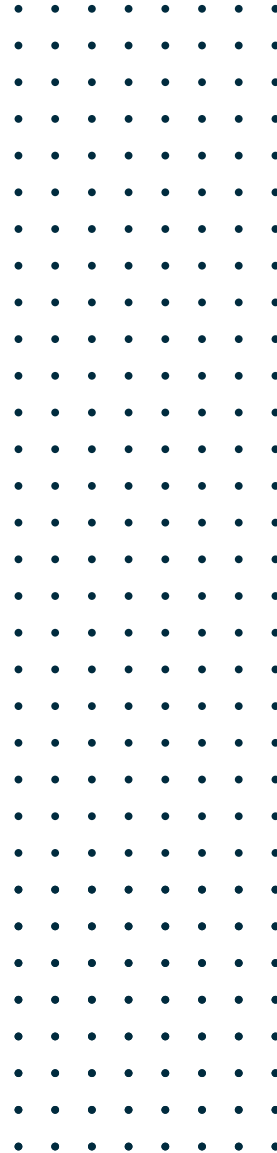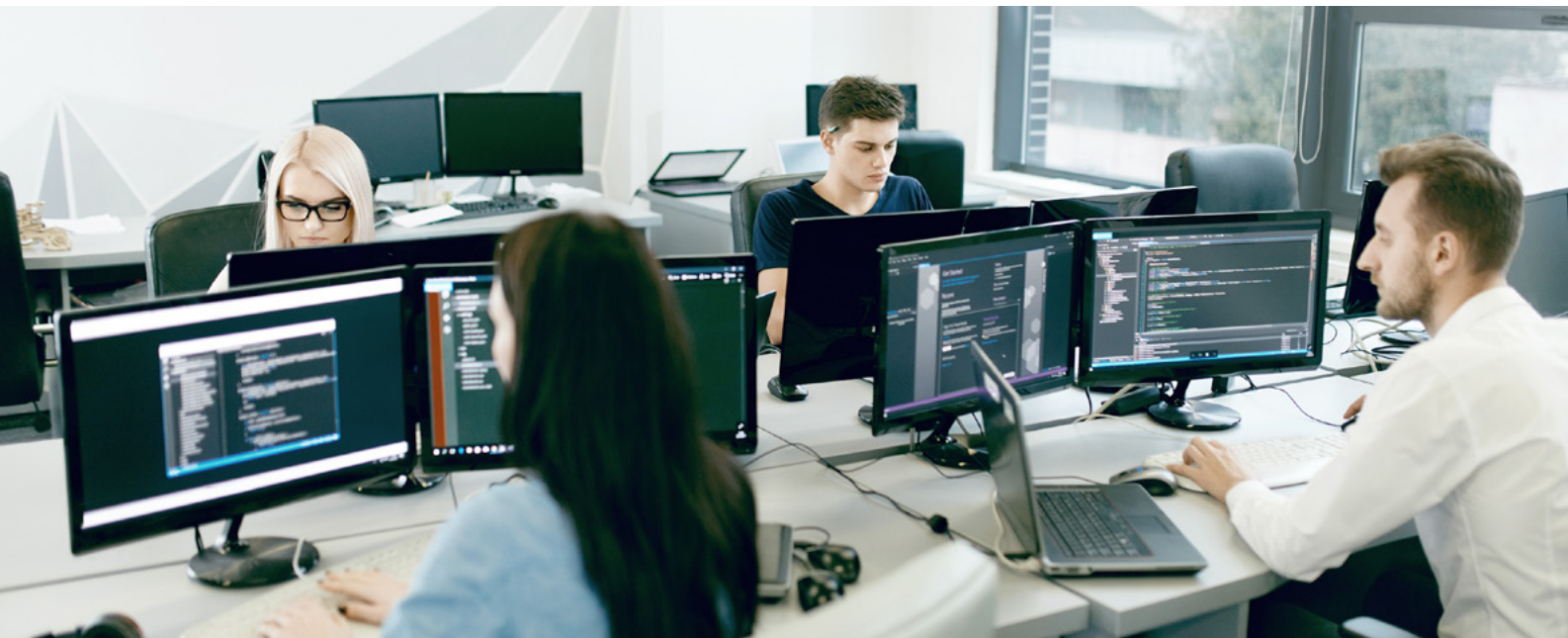| IT	Business User
Data lineage shows the complex variations of the data	Data lineage provides a roadmap of data consistency, which enables business users to trust their data
Data lineage illustrates the physical storage and movement of the data	Data lineage provides enriched context, which helps business users intelligently match the right data sets to the right business purpose.
Unlock the value of your data	

With automated data lineage, in only ten minutes, over ten years of a company's development is documented

– **Martin Masarik**
  Founder of SQLdep
  (now part of Collibra)

Data lineage helps all Data Citizens discover, trust and understand their data. These diagrams make data meaningful by driving data value. They assist IT and the business user in numerous cases such as, compliance and generating business insights. Data lineage helps:
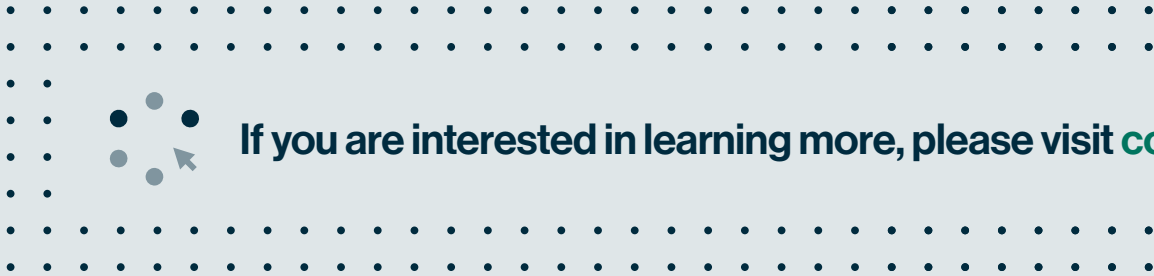
- Companies in the planning and execution of data modernization initiatives (e.g., DWH to cloud) by gaining a complete understanding of the data landscape

- Companies comply with regulations, such as GDPR/CCPA

- Enable impact analysis of any changes on downstream systems

So why should I invest in data lineage?

You may ask yourself, why do I need data lineage? Will it really speed up discovery to insight? Will it really help me make better business decisions? The answer is yes. Data lineage ensures that all Data Citizens **understand** and **trust** their data:

- Data lineage answers the "5 W's" of data. Knowing where your data comes from, who uses it, when it was created, what information it contains and why it exists is crucial for making critical, informed and impactful decisions

- Data lineage provides necessary context and information that helps business users intelligently choose the right data for the right decision

If you are interested in learning more, please visit collibra.com