Business Insights

Harvard Business School Online's Business Insights Blog provides the career insights you need to achieve your goals and gain confidence in your business skills.

5 KEY ELEMENTS OF A DATA ECOSYSTEM

02 MAR 2021

Tim Stobierski | Contributors

Analytics, Data Science Principles, Data Science for Business

When thinking about your business's data, your mind may conjure images of spreadsheets, databases, graphs, and charts. While these are important to your organization's data structure, they're small parts of an extensive data ecosystem.

Whether you're an aspiring data scientist or analyst who wants to directly work with data or a manager who relies on data for decision-making, having a firm understanding of the components that make up your organization's data ecosystem is critical.

Here's an overview of what a data ecosystem is and the key components you should know.

FREE E-BOOK: A BEGINNER'S GUIDE TO DATA & ANALYTICS

Access your free e-book today.

DOWNLOAD NOW >

WHAT IS A DATA ECOSYSTEM?

The term **data ecosystem** refers to the programming languages, packages, algorithms, cloud-computing services, and general infrastructure an organization uses to collect, store, analyze, and leverage data.

No two organizations leverage the same data in the same way. As such, each organization has a unique data ecosystem. These ecosystems may overlap in some cases, particularly when data is pulled or scraped from a public source, or when third-party providers are leveraged (for example, cloud storage providers).

Hi there 👋 Welcome to HBS Online! Can I help you find something?

In the online course Data Science Principles, the concept of the data ecosystem is explored through the lens of key stages in the data project life cycle: sensing, collection, wrangling, analysis, and storage.

COMPONENTS OF A DATA ECOSYSTEM

1. Sensing

Sensing refers to the process of identifying data sources for your project. It involves evaluating the quality of data so you can better understand whether it's valuable. This evaluation includes asking such questions as:

- Is the data accurate?
- Is the data recent and up to date?
- Is the data complete?
- Is the data valid? Can it be trusted?

Data can be sourced from internal sources, such as databases, spreadsheets, CRMs, and other software. It can also be sourced from external sources, such as websites or third-party data aggregators.

Key pieces of the data ecosystem leveraged in this stage include:

- **Internal data sources:** Proprietary databases, spreadsheets, and other resources that originate from within your organization
- **External data sources:** Databases, spreadsheets, websites, and other data sources that originate from outside your organization
- **Software:** Custom software that exists for the sole purpose of data sensing
- **Algorithms:** A set of steps or rules that automates the process of evaluating data for accuracy and completion before it's used

2. Collection

Once a potential data source has been identified, data must be **collected**.

Data collection can be completed through manual or automated processes. That being said, it generally isn't feasible to manually perform large-scale data collection. That's why data scientists use programming languages to write software designed to automate the data collection process.

For example, it's possible to write a piece of code designed to "scrape" relevant information from a website (aptly named a **web scraper**). It's also possible to design and code an **application programming interface**, or **API**, to directly extract specific information from a database or interact with a web application.

Key pieces of the data ecosystem leveraged in this stage include:

- **Various programming languages**: These include R, Python, SQL, and JavaScript
- **Code packages and libraries:** Existing code that's been written and tested and allows data scientists to generate programs more quickly and efficiently
- **APIs:** Software programs designed to interact with other applications and extract data

3. Wrangling

Data wrangling is a set of processes designed to transform raw data into a more usable format. Depending on the quality of the data in question, it may involve merging multiple datasets, identifying and filling gaps in data, deleting unnecessary or incorrect data, and "cleaning" and structuring data for future analysis.

Hi there 👋 Welcome to HBS Online! Can I help you find something?

As with data collection, data wrangling can be performed manually or in an automated fashion. If a dataset is small enough, manual processes can work well. For most larger data projects, the amount of data is too vast and requires automation.

Key pieces of the data ecosystem leveraged in this stage include:

- **Algorithms:** A series of steps or rules to be followed to solve a problem (in this case, the evaluation and manipulation of data)
- **Various programming languages:** These include R, Python, SQL, and JavaScript, and can be used to write algorithms
- **Data wrangling tools:** A variety of data wrangling tools can be purchased or sourced for free to perform parts of the data wrangling process. OpenRefine, DataWrangler, and CSVKit are all examples.

4. Analysis

After raw data has been inspected and transformed into a readily usable state, it can be **analyzed.** Depending on the specific challenge your data project seeks to address, this analysis can be diagnostic, descriptive, predictive, or prescriptive. While each of these forms of analysis is unique, they rely on the same processes and tools.

Typically, your analysis begins with some form of automation, especially when datasets are exceptionally large. After automated processes have been completed, data analysts use their expertise to glean additional insights.

Key pieces of the data ecosystem leveraged in this stage include:

- **Algorithms:** A series of steps or rules to be followed to solve a problem (in this case, the analysis of various data points)
- **Statistical models:** Mathematical models used to investigate and interpret data
- **Data visualization tools:** These include Tableau, Microsoft BI, and Google Charts, which can generate graphical representations of data. Data visualization software may also have other functionality you can leverage.

5. Storage

Throughout all of the data life cycle stages, data must be **stored** in a way that's both secure and accessible. The exact medium used for storage is dictated by your organization's data governance procedures.

Key pieces of the data ecosystem leveraged in this stage include:

- **Cloud-based storage solutions:** These allow an organization to store data off-site and access it remotely
- **On-site servers:** These give organizations a greater sense of control over how data is stored and used
- **Other storage media:** These include hard drives, USB devices, CD-ROMs, and floppy disks

THE IMPORTANCE OF THE DATA ECOSYSTEM

Each component of the data ecosystem interacts with and exerts influence over the other components, which means it can introduce data integrity, privacy, and security threats if a business isn't careful.

For example, consider the recent SolarWinds hack, which has been called one of the worst security breaches in history. SolarWinds is an information technology company that develops software more than 30,000 client companies and organizations use to manage their networks. As such, it's an integral piece of their data ecosystems.

Hi there 👋 Welcome to HBS Online! Can I help you find something?

In early 2020, hackers added malicious code to SolarWinds' software, which was then distributed to clients in the form of updates. As a result, thousands of companies and organizations had their data exposed to hackers, including NASA, the Federal Aviation Administration, and other government agencies.

By understanding how each component of your organization's data ecosystem interacts with other components, you can prepare for these kinds of challenges and identify opportunities for efficiency.

Are you interested in improving your understanding of data science? Learn more about Data Science Principles and Data Science for Business, two online analytics courses designed to help you build your data proficiency.

About the Author

Tim Stobierski is a marketing specialist and contributing writer for Harvard Business School Online.

Hi there 👋 Welcome to HBS Online! Can I help you find something?