Origins of the data analysis process

When you decided to join this program, you proved that you are a curious person. So now, tap into your curiosity and explore the origins of data analysis. No one knows when or why the first person decided to record data about people and things. But it was certainly a smart idea!

Data analysis is rooted in statistics, which has a pretty long history itself. Archaeologists mark the start of statistics in ancient Egypt with the building of the pyramids. The ancient Egyptians were masters of organizing data. They documented their calculations and theories on papyri (paper-like materials), which are now viewed as the earliest examples of spreadsheets and checklists. Today's data analysts owe a lot to those brilliant scribes, who helped create a more technical and efficient process.

It is time to enter the **data analysis process**—how industry professionals move from data to decision. All team members can drive success by planning work both upfront and at the end of the data analysis process. While the data analysis process is well known among experts, there isn't a single defined structure or single architecture that's uniformly followed by every data analyst. But there are some shared fundamentals. This reading provides an overview of several processes, starting with the one that forms the foundation of the Google Data Analytics Certificate.

The process presented as part of the Google Data Analytics Certificate is one that will be valuable to you as you keep moving forward in your career:

1. **Ask**: business challenge, objective, or question
2. **Prepare**: data generation, collection, storage, and data management
3. **Process**: data cleaning and data integrity
4. **Analyze**: data exploration, visualization, and analysis
5. **Share**: communicating and interpreting results
6. **Act**: putting insights to work to solve the problem

Understanding this process—and all of the iterations that helped make it popular—will be a big part of guiding your own analysis and your work in this program. Let's go over a few other variations of the data analysis process.

EMC's data analysis process

EMC Corporation's data analytics process is cyclical with six steps:

1. Discovery
2. Pre-processing data
3. Model planning
4. Model building
5. Communicate results
6. Operationalize

EMC Corporation is now Dell EMC. This model, created by David Dietrich, reflects the cyclical nature of typical business projects. The phases aren't static milestones; each step connects and leads to the next, and eventually repeats. Key questions help analysts test whether they have accomplished enough to move forward and ensure that teams have spent enough time on each of the phases and don't start modeling before the data is ready. It is a little different from the data analysis process on which this program is based on, but it has some core ideas in common: the first phase is interested in discovering and asking questions; data has to be prepared before it can be analyzed and used; and then findings should be shared and acted on.
For more information, refer to this e-book, Data Science & Big Data Analytics.

SAS's iterative process

An iterative data analysis process was created by a company called **SAS**, a leading data analytics solutions provider. It can be used to produce repeatable, reliable, and predictive results:

1. Ask
2. Prepare
3. Explore
4. Model
5. Implement
6. Act
7. Evaluate

The SAS model emphasizes the cyclical nature of their model by visualizing it as an infinity symbol. Its process has seven steps, many of which mirror the other models, like ask, prepare, model, and act. But this process is also a little different; it includes a step after the act phase designed to help analysts evaluate their solutions and potentially return to the ask phase again.

Project-based data analytics process

A project-based data analytics process has five simple steps:

1. Identifying the problem
2. Designing data requirements
3. Pre-processing data
4. Performing data analysis

5. Visualizing data

This data analytics project process was developed by Vignesh Prajapati. It doesn't include the sixth phase, or the act phase. However, it still covers a lot of the same steps described. It begins with identifying the problem, preparing and processing data before analysis, and ends with data visualization.
For more information, refer to Understanding the data analytics project life cycle.

Big data analytics process

Authors Thomas Erl, Wajid Khattak, and Paul Buhler proposed a big data analytics process in their book, **Big Data Fundamentals: Concepts, Drivers & Techniques**. Their process suggests phases divided into nine steps:

1. Business case evaluation
2. Data identification
3. Data acquisition and filtering
4. Data extraction
5. Data validation and cleaning
6. Data aggregation and representation
7. Data analysis
8. Data visualization
9. Utilization of analysis results

This process appears to have three or four more steps than the previous models. But in reality, they have just broken down what has been referred to as prepare and process into smaller steps. It emphasizes the individual tasks required for gathering, preparing, and cleaning data before the analysis phase.
For more information, refer to Big Data Adoption and Planning Considerations.

Key takeaway

From a journey to the pyramids and data in ancient Egypt to now, the way people analyze data has evolved (and continues to do so). The data analysis process is like real life architecture: There are different ways to do things but the same core ideas still appear in each model of the process. Whether you use the structure of this Google Data Analytics Certificate or one of the many other iterations you have learned about, your approach will be effective.

Data and gut instinct

Detectives and data analysts have a lot in common. Both depend on facts and clues to make decisions. Both collect and look at the evidence. Both talk to people who know part of the story. And both might even follow some footprints to see where they lead. Whether you're a detective or a data analyst, your job is all about following steps to collect and understand facts.
Analysts use data-driven decision-making and follow a step-by-step process. You have learned that there are six steps to this process:

1. **Ask** questions and define the problem.
2. **Prepare** data by collecting and storing the information.
3. **Process** data by cleaning and checking the information.
4. **Analyze** data to find patterns, relationships, and trends.
5. **Share** data with your audience.
6. **Act** on the data and use the analysis results.

But there are other factors that influence the decision-making process. You may have read mysteries where the detective used their gut instinct, and followed a hunch that helped them solve the case. **Gut instinct** is an intuitive understanding of something with little or no explanation. This isn't always something conscious; we often pick up on signals without even realizing. You just have a "feeling" it's right.

Why gut instinct can be a problem

At the heart of data-driven decision making is data. Therefore, it's essential that data analysts focus on the data to ensure they make informed decisions. If you ignore data by preferring to make decisions based on your own experience, your decisions may be biased. But even worse, decisions based on gut instinct without any data to back them up can cause mistakes.

Consider an example of a restaurant entrepreneur, partnering with a well-known chef to develop a new restaurant in a bustling part of the city's central shopping district. The chef has several restaurants across the city. Banking on their reputation, the restaurant entrepreneur and chef followed gut instinct and created another uniquely themed restaurant. However, fundraising efforts fell short to fund the opening of the restaurant after months of planning and preparation. The property will go back on the market to be sold at a loss. Had the entrepreneur done more research, they would've found data showing prospective customers in this new restaurant location were very different from the chef's other restaurants.

The more you understand the data related to a project, the easier it will be to figure out what is required. These efforts will also help you identify errors and gaps in your data so you can communicate your findings more effectively. Sometimes past experience helps you make a connection that no one else would notice. For example, a detective might be able to crack open a case because they remember an old case just like the one they're solving today. It's not just gut instinct.

Data + business knowledge = mystery solved

Blending data with business knowledge, plus maybe a touch of gut instinct, will be a common part of your process as a junior data analyst. The key is figuring out the exact mix for each particular project. A lot of times, it will depend on the goals of your analysis. That is why analysts often ask, "How do I define success for this project?"

In addition, try asking yourself these questions about a project to help find the perfect balance:

- What kind of results are needed?
- Who will be informed?
- Am I answering the question being asked?
- How quickly does a decision need to be made?

For instance, if you are working on a rush project, you might need to rely on your own knowledge and experience more than usual. There just isn't enough time to thoroughly analyze all of the available data. But if you get a project that involves plenty of time and resources, then the best strategy is to be more data-driven. It's up to you, the data analyst, to make the best possible choice. You will probably blend data and knowledge a million different ways over the course of your data analytics career. And the more you practice, the better you will get at finding that perfect blend.

Key takeaways

Data analysts and detectives share a similar approach to problem-solving, both relying on evidence and facts to make decisions. Data-driven decision-making is essential for analysts, but gut instinct can also play a role in identifying patterns and connections. Balancing data and gut instinct is crucial for making informed decisions, and the right mix depends on the project's goals and time constraints.