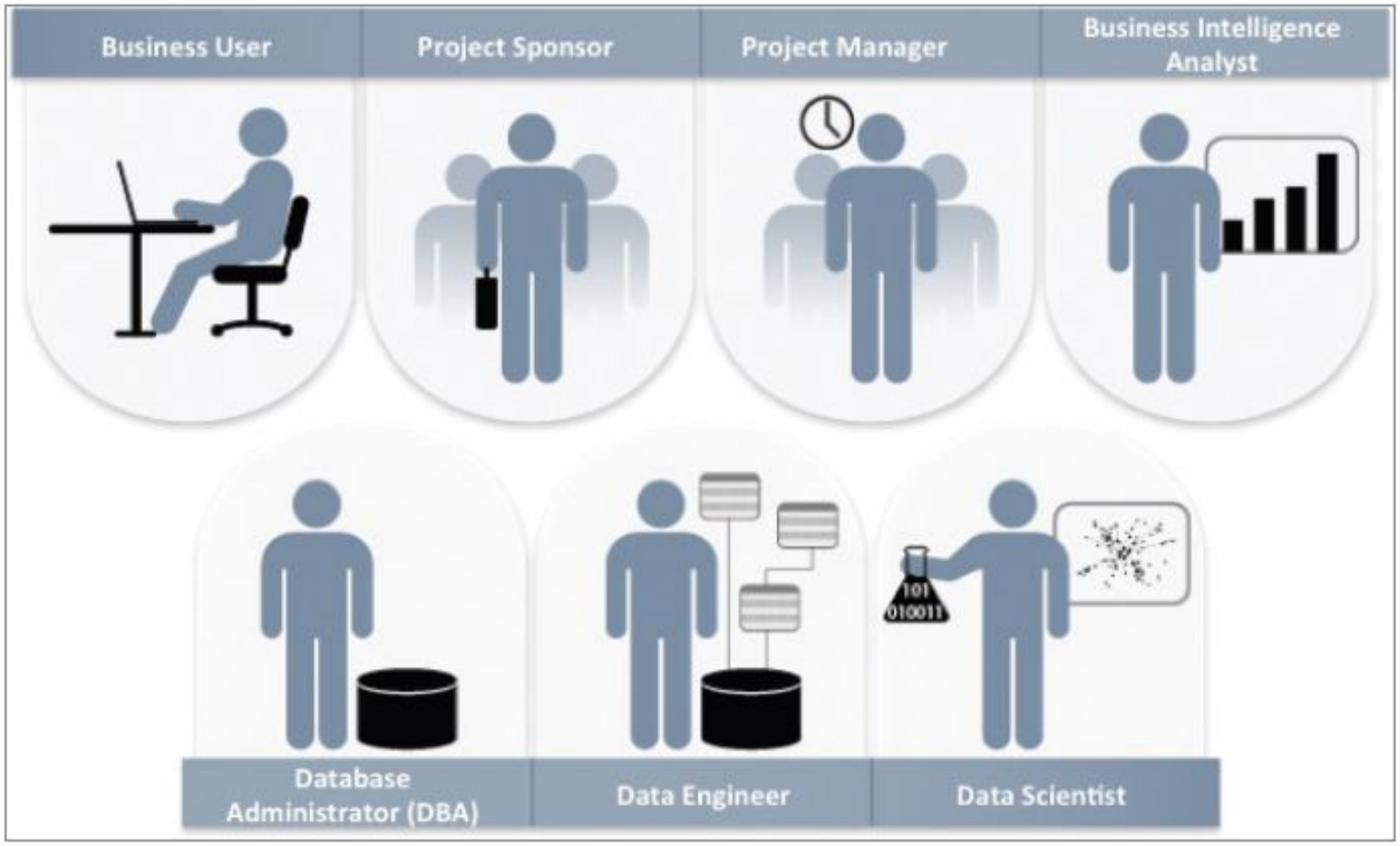
Need of Data Analytics

- Data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. They got value in the following ways:

- **Cost reduction**: Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.
- **Faster, better decision making**: With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned.
- **New products and services:** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

Key Roles for Successful Analytic Projects



Key Roles for Successful Analytic Projects

Business User: Someone who understands the domain area and usually benefits from the results. This person can consult and advise the project team in the context of the project, the value of the results, and how the outputs will be operationalized. Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfills this role.

Project Sponsor: Responsible for the genesis of the project. Provides the impetus and requirements for the project and defines the core business problem. Generally provides the funding and gauges the degree of value from the final outputs of the working team. This person sets the priorities for the project and clarifies the desired outputs.

Project Manager: Ensures that key milestones and objectives are met on time and at the expected quality.

Business Intelligence Analyst: Provides business domain expertise based on a deep understanding of the data, key performance indicators (KPIs), key metrics, and business intelligence from a reporting perspective. Business Intelligence Analysts generally create dashboards and reports and know the data feeds and sources.

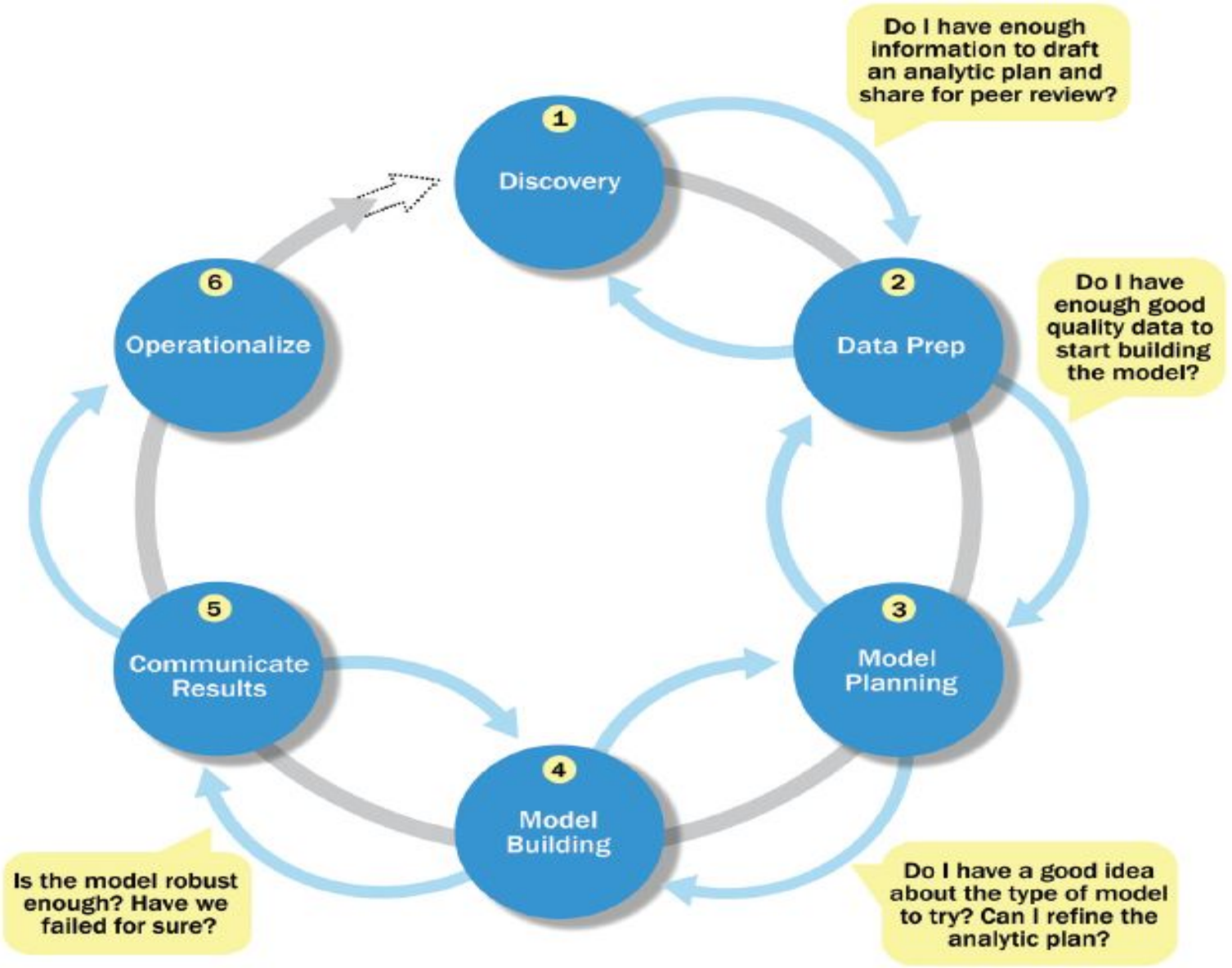
Key Roles for Successful Analytic Projects

- **Database Administrator (DBA):** Provisions and configures the database environment to support the analytics needs of the working team. These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.

- **Data Engineer:** Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox, which was discussed in Chapter 1, "Introduction to Big Data Analytics." Whereas the DBA sets up and configures the databases to be used, the data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics. The data engineer works closely with the data scientist to help shape data in the right ways for analyses.

Data Scientist: Provides subject matter expertise for analytical techniques, data modeling, and applying valid analytical techniques to given business problems. Ensures overall analytics objectives are met. Designs and executes analytical methods and approaches with the data available to the project.

Data Analytics Lifecycle



Phase 1—Discovery:

In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn.

The team assesses the resources available to support the project in terms of people, technology, time, and data.

Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.

Phase 1—Discovery:

1) Learning Business Domain

"Domain knowledge is the knowledge about the environment in which the data is processed to reveal secrets of the data".

In other words, the knowledge of the field that the data belongs to is known as Domain Knowledge.

- For example, the knowledge of the automobile industry when working with the relevant data can be used like — Let's say we have two features Horsepower and RPM from which we can create an additional feature like Torque from the formula
-

  TORQUE = HP x 5252 ÷ RPM

- Feature engineering is creating features using the domain knowledge to optimize the machine learning algorithms.
- The earlier the team can make this assessment the better, because the decision helps dictate the resources needed for the project team and ensures the team has the right balance of domain knowledge and technical expertise.

Phase 1—Discovery:

2) Resources:

The team needs to assess the resources available to support the project. In this context, resources include technology, tools, systems, data, and people.

Tools and Technology:

a) What are the available systems, tools and technology?

b) What other tools will be required In later phases of the project?

Skills:

a)What types of skills and roles will be required that may not exist today?

b)For the project to have long-term success, what types of skills and roles will be needed for the recipients of the model being developed?

c)Does the requisite level of expertise exist within the organization today, or will it need to be cultivated?

Data:

a) data available is sufficient to support the project's goals.?

b) Is it required to collect additional data, purchase it from outside sources, or transform existing data?

c) When the data is less than hoped for, the size and scope of the project is reduced to work within the constraints of the existing data.

Phase 1—Discovery:

3) Framing the right problem:

Framing is the process of stating the analytics problem to be solved.

It is important to identify the

- main objectives of the project,
- what needs to be achieved in business terms,
- what needs to be done to meet the needs.

Additionally, consider the objectives and the success criteria for the project.

What is the team attempting to achieve by doing the project, and what will be considered "good enough" as an outcome of the project?

It is best practice to share the statement of goals and success criteria with the team and confirm alignment with the project sponsor's expectations.

Equally important is to establish failure criteria. The failure criteria will guide the team in understanding when it is best to stop trying or settle for the results that have been gleaned from the data.

Establishing criteria for both success and failure helps the participants avoid unproductive effort and remain aligned with the project sponsors

Phase 1—Discovery:

4) Identifying Key Stakeholders: Another important step is to identify the key stakeholders, which should include anyone who will benefit from the project or will be significantly impacted by the project and their interests in the project.

Depending on the number of stakeholders and participants, the team may consider outlining the type of activity and participation expected from each stakeholder and participant.

This will set clear expectations with the participants and avoid delays later when, for example, theteam may feel it needs to wait for approval from someone who views himself as an adviser rather than an approver of the work product.

When interviewing stakeholders, learn about the domain area and any relevant history from similar analytics projects. For example, the team may identify the results each stakeholder wants from the project and the criteria it will use to judge the success of the project.

5) Interviewing the Analytics Sponsor

Following is a brief list of common questions that are helpful to ask during the discovery phase when interviewing the project sponsor. The responses will begin to shape the scope of the project and give the team an idea of the goals and objectives of the project.

- What business problem is the team trying to solve?
- What is the desired outcome of the project?
- What data sources are available?
- What industry issues may impact the analysis?
- What timelines need to be considered?
- Who could provide insight into the project?
- Who has final decision-making authority on the project?
- How will the focus and scope of the problem change if the following dimensions change:
- Time: Analyzing 1 year or 10 years' worth of data?
- People: Assess impact of changes in resources on project timeline.
- Risk: Conservative to aggressive
- Resources: None to unlimited (tools, technology, systems)
- Size and attributes of data: Including internal and external data sources

Phase 1—Discovery:

6) Developing Initial Hypotheses

This step involves forming ideas that the team can test with data.

Generally, it is best to come up with a few primary hypotheses to test and then be creative about developing several more.

In this way, the team can compare its answers with the outcome of an experiment or test to generate additional possible solutions to problems. As a result, the team will have a much richer set of observations to choose from and more choices for agreeing upon the most impactful conclusions from a project.

Another part of this process involves gathering and assessing hypotheses from stakeholders and domain experts who may have their own perspective on what the problem is, what the solution should be, and how to arrive at a solution. These stakeholders would know the domain area well and can offer suggestions on

ideas to test as the team formulates hypotheses during this phase. The team will likely collect many ideas that may illuminate the operating assumptions of the stakeholders.

7) Identifying Potential Data Sources

Consider the volume, type, and time span of the data needed to test the hypotheses. Ensure that the team can access more than simply aggregated data.

The team should perform five main activities during this step of the discovery phase:

- Identify data sources: Make a list of candidate data sources the team may need to test the initial hypotheses outlined in this phase. Make an inventory of the datasets currently available and those that can be purchased or otherwise acquired for the tests the team wants to perform.
- Capture aggregate data sources: This is for previewing the data and providing high-level understanding. It enables the team to gain a quick overview of the data and perform further exploration on specific areas. It also points the team to possible areas of interest within the data.
- Review the raw data: Obtain preliminary data from initial data feeds. Begin understanding the interdependencies among the data attributes, and become familiar with the content of the data, its quality, and its limitations.
- Evaluate the data structures and tools needed: The data type and structure dictate which tools the team can use to analyze the data. This evaluation gets the team thinking about which technologies may be good candidates for the project and how to start getting access to these tools.
- Scope the sort of data infrastructure needed for this type of problem: In addition to the tools needed, the data influences the kind of infrastructure that's required, such as disk storage and network capacity.

Data Preparation

Phase 2—Data preparation: Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT. Data should be transformed in the ETLT process so the team can work with it and analyze it. In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data..

Data Preparation

- Data preparation, which includes the steps to explore, pre process, and condition data prior to modelling and analysis.
- 1) **Preparing the Analytic Sandbox:** The first subphase of data preparation requires the team to obtain an analytic sandbox (also commonly referred to as a ***workspace), in which the team can explore the data without interfering with live production*** databases. Consider an example in which the team needs to work with a company's financial data. The team should access a copy of the financial data from the analytic sandbox rather than interacting with the production version of the organization's main database, because that will be tightly controlled and needed for financial reporting.
- **2) Performing ELT/ETL:**
- In ETL, users perform extract, transform, load processes to extract data from a datastore, perform data transformations, and load the data back into the datastore.
- In ELT, the data is extracted in its raw form and loaded into the datastore, where analysts can choose to transform the data into a new state or leave it in its original, raw condition.

- Depending on the size and number of the data sources, the team may need to consider how to parallelize the movement of the datasets into the sandbox.

Data Preparation

3) Learning About the Data

- A critical aspect of a data science project is to become familiar with the data itself. Spending time to learn the nuances of the datasets provides context to understand what constitutes a reasonable value and expected output versus what is a surprising finding. In addition, it is important to catalog the data sources that the team has access to and identify additional data sources that the team can leverage but perhaps does not have access to today. Some of the activities in this step may overlap with the initial investigation of the datasets that occur in the discovery phase.

4)**Data Conditioning**

- ***Data conditioning refers to the process of cleaning data, normalizing datasets, and performing transformations*** on the data. A critical step within the Data Analytics Lifecycle, data conditioning can involve many complex steps to join or merge datasets or otherwise get datasets into a state that enables analysis in further phases. Data conditioning is often viewed as a pre processing step for the data analysis because it involves many operations on the dataset before developing models to process or analyze the data. This implies that the data-conditioning step is performed only by IT, the data owners, a DBA, or a data engineer.

Data Preparation

5) **Survey and Visualize:**

Seeing high-level patterns in the data enables one to understand characteristics about the data very quickly. One example is using data visualization to examine data quality, such as whether the data contains many unexpected values or other indicators of dirty data. Another example is skewness, such as if the majority of the data is heavily shifted toward one value or end of a continuum.

a) Does the data distribution stay consistent over all the data? If not, what kinds of actions should be taken to address this problem?

b) Assess the granularity of the data, the range of values, and the level of aggregation of the data.

c) Does the data represent the population of interest? For marketing data, if the project is focused on targeting customers of child-rearing age, does the data represent that, or is it full of senior citizens and teenagers?

d) For time-related variables, are the measurements daily, weekly, monthly? Is that good enough? Is time measured in seconds everywhere? Or is it in milliseconds in some places? Determine the level of granularity of the data needed for the analysis, and assess whether the current level of timestamps on the data meets that need.

e) Is the data standardized/normalized? Are the scales consistent? If not, how consistent or irregular is the data?

f) For geospatial datasets, are state or country abbreviations consistent across the data? Are personal names normalized? English units? Metric units?

Data Preparation

Common Tools for the Data Preparation Phase:

Hadoop can perform massively parallel ingest and custom analysis for web traffic parsing, GPS location analytics, genomic analysis, and combining of massive unstructured data feeds from multiple sources.

- Alpine Miner provides a graphical user interface (GUI) for creating analytic workflows, including data manipulations and a series of analytic events such as staged data-mining techniques (for example, first select the top 100 customers, and then run descriptive statistics and clustering) on Postgres SQL and other Big Data sources.
- OpenRefine (formerly called Google Refine) is "a free, open source, powerful tool for working with messy data." It is a popular GUI-based tool for performing data transformations, and it's one of the most robust free tools currently available.
- Similar to OpenRefine, Data Wrangler is an interactive tool for data cleaning and transformation.

Wrangler was developed at Stanford University and can be used to perform many transformations on a given dataset. In addition, data transformation outputs can be put into Java or Python. The advantage of this feature is that a subset of the data can be manipulated in Wrangler via its GUI, and then the same operations can be written out as Java or Python code to be executed against the full, larger dataset offline in a local analytic sandbox.

Phase 3 Model Planning

Phase 3—Model planning: Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

1) Data Exploration and Variable Selection:
The objective of the data exploration is to understand the relationships among the variables to inform selection of the variables and methods and to understand the problem domain.

2) Model Selection: In the model selection subphase, the team's main goal is to choose an analytical technique, or a short list of candidate techniques, based on the end goal of the project.

Model Planning

Common Tools for the Model Planning Phase

Many tools are available to assist in this phase. Here are several of the more common ones:

- R has a complete set of modeling capabilities and provides a good environment for building interpretive models with high-quality code.
- SQL Analysis services can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models.
- SAS/ACCESS provides integration between SAS and the analytics sandbox via multiple data connectors such as OBDC, JDBC, and OLE DB. SAS itself is generally used on file extracts, but with SAS/ACCESS, users can connect to relational databases (such as Oracle or Teradata) and data warehouse appliances (such as Greenplum or Aster), files, and enterprise applications (such as SAP and Salesforce.com).

Model building

Phase 4—Model building: In Phase 4, the team develops datasets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing, if applicable).

Model Building

- Data science team needs to develop datasets for training, testing, and production purposes.
- These datasets enable the data scientist to develop the analytical model and train it (“training data”), while holding aside some of the data (“hold-out data” or “test data”) for testing the model. During this process, it is critical to ensure that the training and test datasets are sufficiently robust for the model and analytical techniques. A simple way to think of these datasets is to view the training dataset for conducting the initial experiments and the test sets for validating an approach once the initial experiments and models have been run.
- In the model building phase, an analytical model is developed and fit on the training data and evaluated (scored) against the test data. The phases of model planning and model building can overlap quite a bit, and in practice one can iterate back and forth between the two phases for a while before settling on a final model.

Model Building

Questions to consider include these:

- Does the model appear valid and accurate on the test data?
- Does the model output/ behavior make sense to the domain experts? That is, does it appear as if the model is giving answers that make sense in this context?
- Do the parameter values of the fitted model make sense in the context of the domain?
- Is the model sufficiently accurate to meet the goal?
- Does the model avoid intolerable mistakes? Depending on context, false positives may be more serious or less serious than false negatives, for instance.
- Are more data or more inputs needed? Do any of the inputs need to be transformed or eliminated?
- Will the kind of model chosen support the runtime requirements?
- Is a different form of the model required to address the business problem? If so, go back to the model planning phase and revise the modeling approach.

Model building

Common Tools for the Model Building Phase

There are many tools available to assist in this phase, focused primarily on statistical analysis or data mining software. Common tools in this space include, but are not limited to, the following:

Commercial Tools:

- SAS Enterprise Miner allows users to run predictive and descriptive models based on large volumes of data from across the enterprise. It interoperates with other large data stores, has many partnerships, and is built for enterprise-level computing and analytics.

- SPSS Modeler (provided by IBM and now called IBM SPSS Modeler) offers methods to explore and analyze data through a GUI.

- Matlab provides a high-level language for performing a variety of data analytics, algorithms, and data exploration.

- Alpine Miner provides a GUI front end for users to develop analytic workflows and interact with Big Data tools and platforms on the back end.

- STATISTICA and Mathematica are also popular and well-regarded data mining and analytics tools.

Model Building

● Free or Open Source tools:

- ● R and PL/R: R was described earlier in the model planning phase, and PL/R is a procedural language for PostgreSQL with R. Using this approach means that R commands can be executed in database. This technique provides higher performance and is more scalable than running R in memory.
- ● Octave : a free software programming language for computational modeling, has some of the functionality of Matlab. Because it is freely available, Octave is used in major universities when teaching machine learning.
- ● WEKA : is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.
- ● Python is a programming language that provides toolkits for machine learning and analysis, such as scikit-learn, numpy, scipy, pandas, and related data visualization using matplotlib.
- ● SQL in-database implementations, such as MADlib, provide an alterative to in-memory desktop analytical tools. MADlib provides an open-source machine learning library of algorithms that can be executed in-database, for PostgreSQL or Greenplum.

Communicate Results

Phase 5—Communicate results: In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

As a result of this phase, the team will have documented the key findings and major insights derived from the analysis. The deliverable of this phase will be the most visible portion of the process to the outside stakeholders and sponsors, so take care to clearly articulate the results, methodology, and business value of the findings.

Operationalize

Phase 6—Operationalize: In Phase 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.
During the pilot project, the team may need to consider executing the algorithm in the database rather than with in-memory tools such as R because the run time is significantly faster and more efficient than running in-memory, especially on larger datasets.

Part of the operationalizing phase includes creating a mechanism for performing ongoing monitoring of model accuracy and, if accuracy degrades, finding ways to retrain the model. If feasible, design alerts for when the model is operating "out-of-bounds." This includes situations when the inputs are beyond the range that the model was trained on, which may cause the outputs of the model to be inaccurate or invalid. If this begins to happen regularly, the model needs to be retrained on new data.

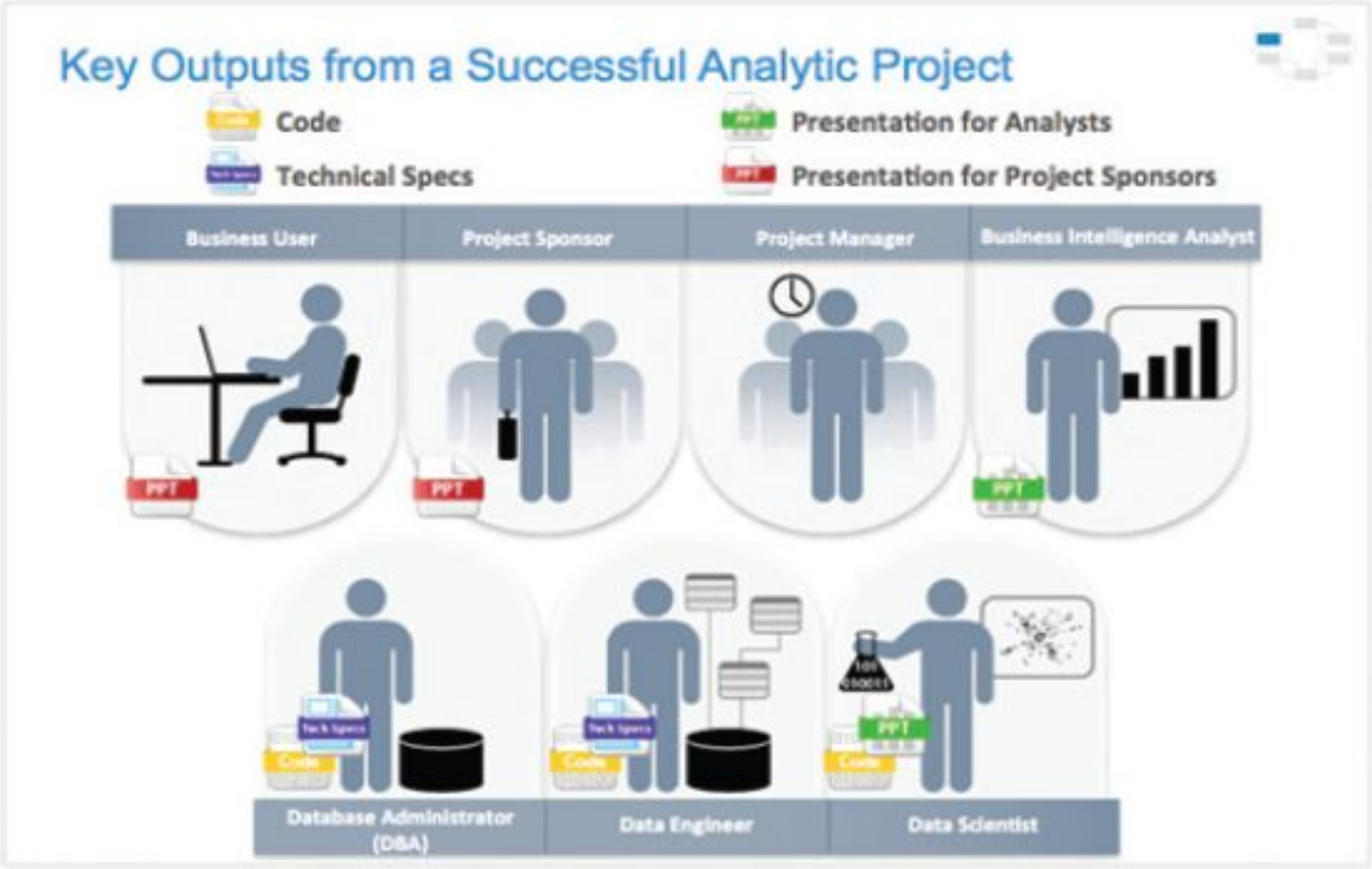
Output for Each Stakeholder

Business User typically tries to determine the benefits and implications of the findings to the business.

- Project Sponsor typically asks questions related to the business impact of the project, the risks and return on investment (ROI), and the way the project can be evangelized within the organization (and beyond).
- Project Manager needs to determine if the project was completed on time and within budget and how well the goals were met.
- Business Intelligence Analyst needs to know if the reports and dashboards he manages will be impacted and need to change.
- Data Engineer and Database Administrator (DBA) typically need to share their code from the analytics project and create a technical document on how to implement it.

Data Scientist needs to share the code and explain the model to her peers, managers, and other stakeholders.

Key Outputs from a Successful Analytics Project



Key Outputs from a Successful Analytics Project

Although these seven roles represent many interests within a project, these interests usually overlap, and most of them can be met with four main deliverables.

- Presentation for project sponsors: This contains high-level takeaways for executive level stakeholders, with a few key messages to aid their decision-making process. Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.
- Presentation for analysts, which describes business process changes and reporting changes. Fellow data scientists will want the details and are comfortable with technical graphs (such as Receiver Operating Characteristic [ROC] curves, density plots, and histograms.
- Code for technical people.
- Technical specifications of implementing the code.