Last Updated: 25 Sep, 2018

How to Structure a Data Science Team: Key Models and Roles to Consider

Share:

Reading time: 12 minutes

If you've been following the direction of expert opinion in data science and predictive analytics, you've likely come across the resolute recommendation to embark on machine learning. As James Hodson in Harvard Business Review recommends, the smartest move is to reach for the "low hanging fruit" and then scale for expertise in heavier operations.

Just recently we talked about machine-learning-as-a-service (MLaaS) platforms. The main takeaway from the current trends is simple. Machine learning becomes more approachable for midsize and small businesses as it gradually turns into a commodity. The leading vendors – Google, Amazon, Microsoft, and IBM – provide APIs and platforms to run basic ML operations without a private infrastructure and deep data science expertise. In the early stages, taking this lean and frugal approach would be the smartest move. As analytics capabilities scale, a team structure can be reshaped to boost operational speed and extend an analytics arsenal.

How to implement this incremental approach? This time we talk about data science team structures and their complexity. Use the following links to jump right to the section you need.

Data science team structures

Embarking on data science and predictive analytics requires a clear understanding of how the initiative is going to be introduced, maintained, and further scaled in terms of team structure. We recommend considering three basic team structures that match different stages of machine learning adoption.

IT-centric structure

Sometimes, hiring data scientists is not an option, and you must leverage talent that's already in-house. The main analytics and leadership role would be a "business translator," usually referred to as a chief analytics officer (CAO) or chief data officer (CDO). The latter term gradually becomes redundant as most data processes are reshaped towards predictive analytics. This person should be capable of leading the initiative. We'll take a more detailed look at the position below.

All the rest – data preparation, training models, creating user interfaces, and model deployment within a corporate IT infrastructure – can be largely managed by the IT department (if your organization actually has a fully functioning, in-house IT department). This approach is fairly limited, but it can be realized by using MLaaS solutions. Environments like Azure Machine Learning or Amazon Machine Learning are already equipped with approachable user interfaces to clean datasets, train models, evaluate them, and deploy.

Azure Machine Learning, for instance, supports its users with detailed documentation for a low entry threshold. This allows for fast training and early deployment of models even without an expert data scientist on board.

On the other hand, MLaaS solutions present their limitations in terms of machine learning methods and cost. All operations, from data cleaning to model evaluation, have their separate prices. And considering that the number of iterations to train an effective model can't be estimated in advance, working with MLaaS platforms entails some budget uncertainty.

Pros of IT-centric structure:

- Leverage new investments with existing IT resources
- Computing infrastructure is provided and maintained by an external service
- In-house specialists can be trained to further realize predictive analytics potential
- Cross-silo management is reduced as all operations are held within the IT department
- Less time-to-market for relatively simple machine learning tasks requiring one or a few models

Cons of IT-centric structure:

- Limited machine learning methods and data cleaning procedures that these services provide

- Model training, testing, and prediction should be paid for. This entails uncertainty of eventual cost per prediction as the number of needed iterations can't be estimated in advance

Integrated structure

With the integrated structure, a data science team focuses on dataset preparation and model training, while IT specialists take charge of the interfaces and infrastructure supporting deployed models. Combining machine learning expertise with IT resource is the most viable option for constant and scalable machine learning operations.

Unlike the IT-centric approach, the integrated method requires having an experienced data scientist on a team and an elaborate recruitment effort beforehand. This ensures better operational flexibility in terms of available techniques. Besides end-to-end and yet limited services, you can leverage deeper machine learning tools and libraries – like Tensor Flow or Theano – that are designed for researchers and experts with data science backgrounds. With this effort allocation, you can address highly specific business problems and choose between as-a-service and custom-built ML solutions.

Pros of integrated structure:

- Leveraging existing IT resources and investments
- Data scientists focus on innovation
- Utilizing full potential of both as-a-service and custom ML applications
- Start with one or two data scientists, then train and onboard more homegrown experts
- Using custom model combinations (ensemble models) that yield better or broader predictions

Cons of integrated structure:

- Computing infrastructure is required in case of custom ML use
- Cross-silo management takes considerable effort
- Significant investments into data science talent acquisition
- Data science talent engagement and retention challenges

Specialized data science department

To reduce management effort and build an all-encompassing machine learning framework, you can run the entire machine learning workflow within an independent data science department. This approach entails the highest cost. All operations, from data cleaning and model training to building front-end interfaces, are realized by a dedicated data science team. It doesn't necessarily mean that all team members should have a data science background, but they should acquire technology infrastructure and service management skills.

A specialized structure model aids in addressing complex data science tasks that include research, use of multiple ML models tailored to various aspects of decision-making, or multiple ML-backed services. In the case of large organizations, specialized data science teams can supplement different business units and operate within their specific fields of analytical interest.

Most successful AI-driven companies operate with specialized data science teams. Obviously, being custom-built and wired for specific tasks, they're all very different. The team structure at Airbnb Data Science is one of the most interesting ones. You can watch this fascinating talk by Airbnb's data scientist Martin Daniel for a deeper understanding of how the company builds its culture or read a blog post from its ex-DS lead, but in short, here are the main principles they apply:

Experiment. Find ways to put data into new projects using an established Learn-Plan-Test-Measure process.

Democratize data. Scale your data science team to the whole company and even clients.

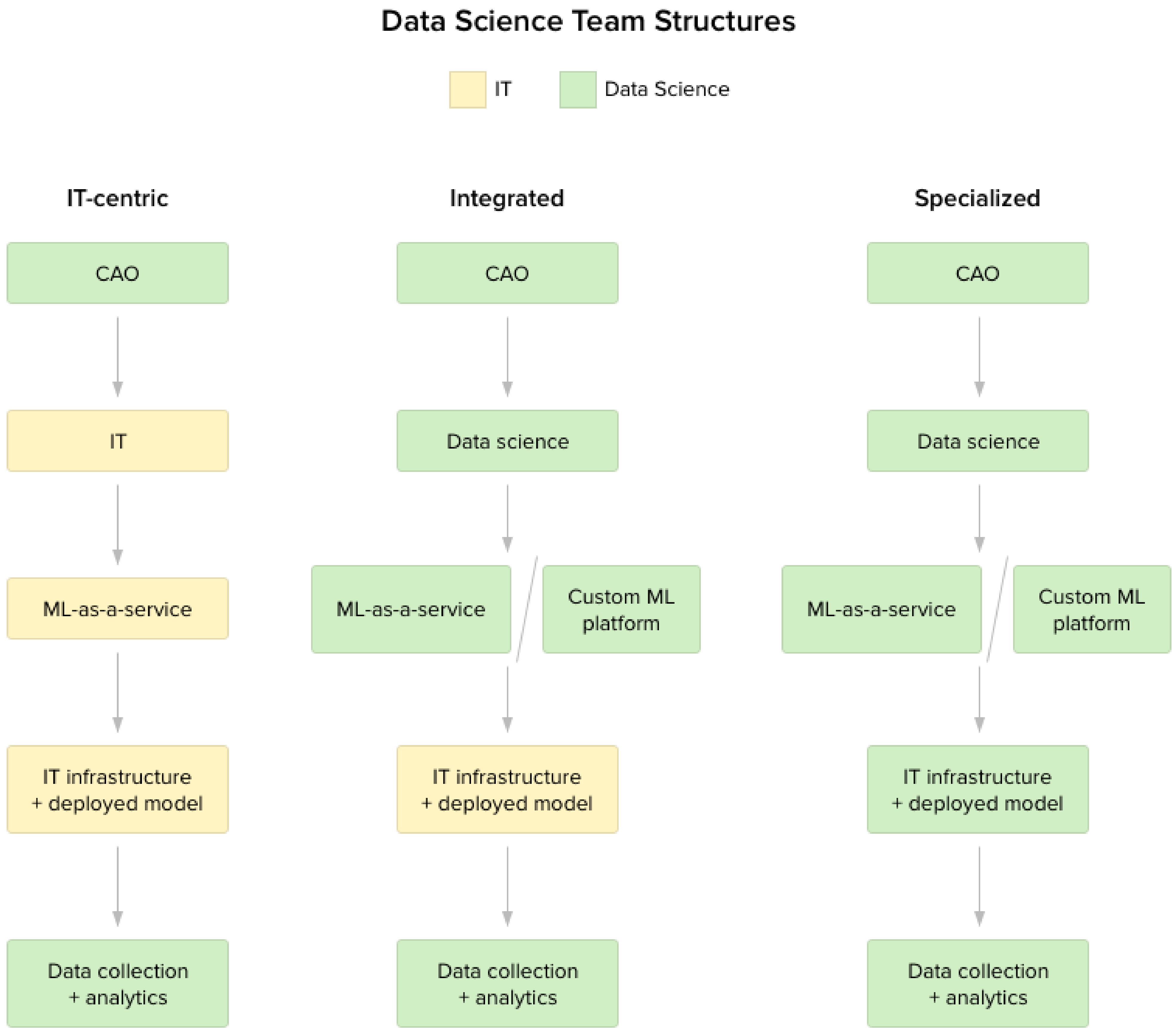
Measure the impact. Evaluate what part DS teams have in your decision-making process and give them credit for it.

Pros of specialized data science department:

- Centralized data science management and increased problem-solving capacities
- Realizing the full potential of both as-a-service and custom ML applications
- Solving complex prediction problems that require deep research or building segmented model factories (that operate automatically across different segments and business units)
- Setting a fully-featured data science playground to foster innovation
- Greater scalability potential

Cons of specialized data science department:

- Building and maintaining a complex computational infrastructure
- Heavy investments into data science talent acquisition
- Data science talent engagement and retention challenges


Data Science Team Structures
IT
Data Science
IT-centric
Integrated
Specialized
CAO
IT
ML-as-a-service
IT infrastructure + deployed model
Data collection + analytics
CAO
Data science
ML-as-a-service
Custom ML platform
IT infrastructure + deployed model
Data collection + analytics
CAO
Data science
ML-as-a-service
Custom ML platform
IT infrastructure + deployed model
Data collection + analytics

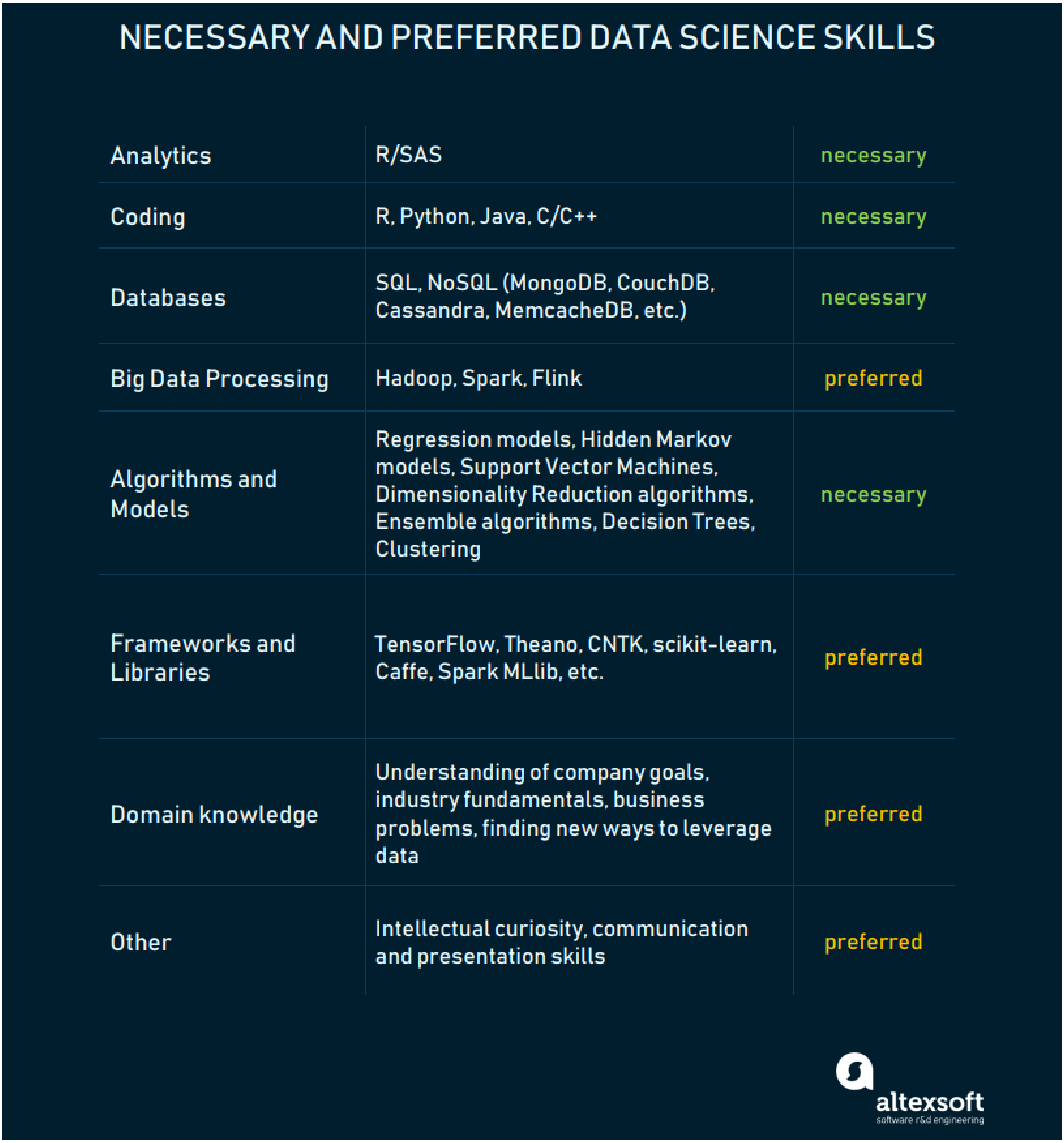
Enterprise IT involvement changes depending on the team structure you choose

Regardless of what structure you opt for to start building data science teams, having the right talent is critical. Who are the people you should look for?

Data science team roles

Let's talk about data scientist skill sets. Unfortunately, the term *data scientist* expanded and became too vague in recent years. After data science appeared in the business spotlight, there is no consensus developed regarding what the skillset of a data scientist is. Matthew Mayo, Data Scientist and the Deputy Editor of KDNuggets, argues: "*When I hear the term **data scientist**, I tend to think of the unicorn, and all that it entails, and then remember that they don't exist, and that actual data scientists play many diverse roles in organizations, with varying levels of business, technical, interpersonal, communication, and domain skills.*"

This is true. It's hard to find unicorns, but it's possible to grow them from people with niche expertise in data science. We at AltexSoft consider these data science skills when hiring machine learning specialists:



NECESSARY AND PREFERRED DATA SCIENCE SKILLS

Analytics	R/SAS	necessary
Coding	R, Python, Java, C/C++	necessary
Databases	SQL, NoSQL (MongoDB, CouchDB, Cassandra, MemcacheDB, etc.)	necessary
Big Data Processing	Hadoop, Spark, Flink	preferred
Algorithms and Models	Regression models, Hidden Markov models, Support Vector Machines, Dimensionality Reduction algorithms, Ensemble algorithms, Decision Trees, Clustering	necessary
Frameworks and Libraries	TensorFlow, Theano, CNTK, scikit-learn, Caffe, Spark MLlib, etc.	preferred
Domain knowledge	Understanding of company goals, industry fundamentals, business problems, finding new ways to leverage data	preferred
Other	Intellectual curiosity, communication and presentation skills	preferred

Skillset of a data scientist

As you will see below, there are many roles within the data science ecosystem, and a lot of classifications offered on the web. We will share with you the one offered by Stitch Fix's Michael Hochster. Michael defines two types of data scientists: Type A and Type B.

Type A stands for Analysis. This person is a statistician that makes sense of data without necessarily having strong programming knowledge. Type A data scientists perform data cleaning, forecasting, modeling, visualization, etc.

Type B stands for Building. These folks use data in production. They're excellent good software engineers with some stats background who build recommendation systems, personalization use cases, etc.

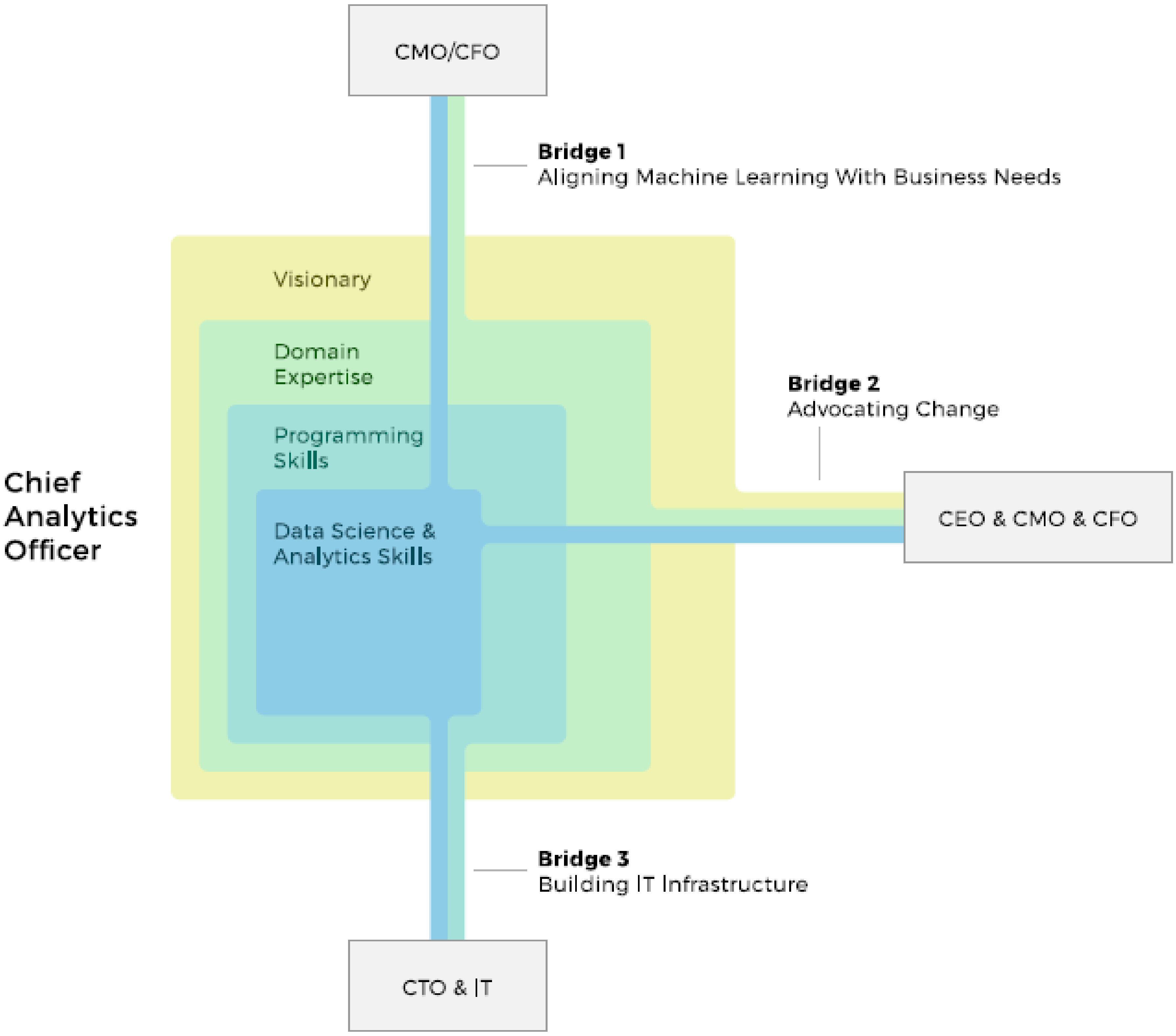
Rarely does one expert fit into a single category. But understanding these two data science functions can help you make sense of the roles we've described further.

Keep in mind that even professionals with this hypothetical skillset usually have their core strengths, which should be considered when distributing roles within a team. In most cases, acquiring talents will entail further training depending on their background.

But people and their roles are two different things. For instance, if your team model is the integrated one, an individual may combine multiple roles. So, let's disregard how many actual experts you may have and outline the roles themselves. Obviously, many skillsets across roles may intersect.

Chief Analytics Officer/Chief Data Officer. In our whitepaper on machine learning, we broadly discussed this key leadership role. CAO, a "business translator," bridges the gap between data science and domain expertise acting both as a visionary and a technical lead. You may get a better idea by looking the visualization below.

Preferred skills: *data science and analytics, programming skills, domain expertise, leadership and visionary abilities*



Role of a Chief Analytics Officer

Data analyst. The data analyst role implies proper data collection and interpretation activities. An analyst ensures that collected data is relevant and exhaustive while also interpreting the analytics results. Some companies, like IBM or HP, also require data analysts to have visualization skills to convert alienating numbers into tangible insights through graphics.

Preferred skills: *R, Python, JavaScript, C/C++, SQL*

Business analyst. A business analyst basically realizes a CAO's functions but on the operational level. This implies converting business expectations into data analysis. If your core data scientist lacks domain expertise, a business analyst bridges this gulf.

Preferred skills: *data visualization, business intelligence, SQL*

Data scientist (not a data science *unicorn*)**.** What does a data scientist do? Assuming you aren't hunting unicorns, a data scientist is a person who solves business tasks using machine learning and data mining techniques. If this is too fuzzy, the role can be narrowed down to data preparation and cleaning with further model training and evaluation.

Preferred skills: *R, SAS, Python, Matlab, SQL, noSQL, Hive, Pig, Hadoop, Spark*

To avoid confusion and make the search for a data scientist less overwhelming, their job is often divided into two roles: machine learning engineer and data journalist.

A **machine learning engineer** combines software engineering and modeling skills by determining which model to use and what data should be used for each model. Probability and statistics are also their forte. Everything that goes into training, monitoring, and maintaining a model is ML engineer's job.

Preferred skills: *R, Python, Scala, Julia, Java*

Data journalists help make sense of data output by putting it in the right context. They're also tasked with articulating business problems and shaping analytics results into compelling stories. Though required to have coding and statistics experience, they should be able to present the idea to stakeholders and represent the data team with those unfamiliar with statistics.

Preferred skills: *SQL, Python, R, Scala, Carto, D3, QGIS, Tableau*

Data architect. This role is critical for working with large amounts of data (you guessed it, Big Data). However, if you don't solely rely on MLaaS cloud platforms, this role is critical to warehouse the data, define database architecture, centralize data, and ensure integrity across different sources. For large distributed systems and big datasets, the architect is also in charge of performance.

Preferred skills: *SQL, noSQL, XML, Hive, Pig, Hadoop, Spark*

Data engineer. Engineers implement, test, and maintain infrastructural components that data architects design. Realistically, the role of an engineer and the role of an architect can be combined in one person. The set of skills is very close.

Preferred skills: *SQL, noSQL, Hive, Pig, Matlab, SAS, Python, Java, Ruby, C++, Perl*

Application/data visualization engineer. Basically, this role is only necessary for a specialized data science model. In other cases, software engineers come from IT units to deliver data science results in applications that end-users face. And it's very likely that an application engineer or other developers from front-end units will oversee end-user data visualization.

Preferred skills: *programming, JavaScript (for visualization), SQL, noSQL*

Team assembly and scaling

The initial challenge of talent acquisition in data science, besides the overall scarcity of experts, is the high salary expectations. According to O'Reilly Data Science Salary Survey 2017, the median annual base salary is $90,000, while in the US the figure reaches $112,000 (up 6.5 percent over last year). These numbers significantly vary depending on geography, specific technology skills, organization sizes, gender, industry, and education. If you decide to hire skilled analytics experts, the further challenges also include engagement and retention.

The intellectual curiosity in combination with the high demand challenges organizations to engage data scientists with creative and explorative projects. Due to these reasons, the *IT-centric team structure* – which leverages existing sources – is a promising alternative on the initial levels of machine learning adoption. Thus, engineers can acquire some analytics skills through ML-as-a-service solutions with friendly interfaces.

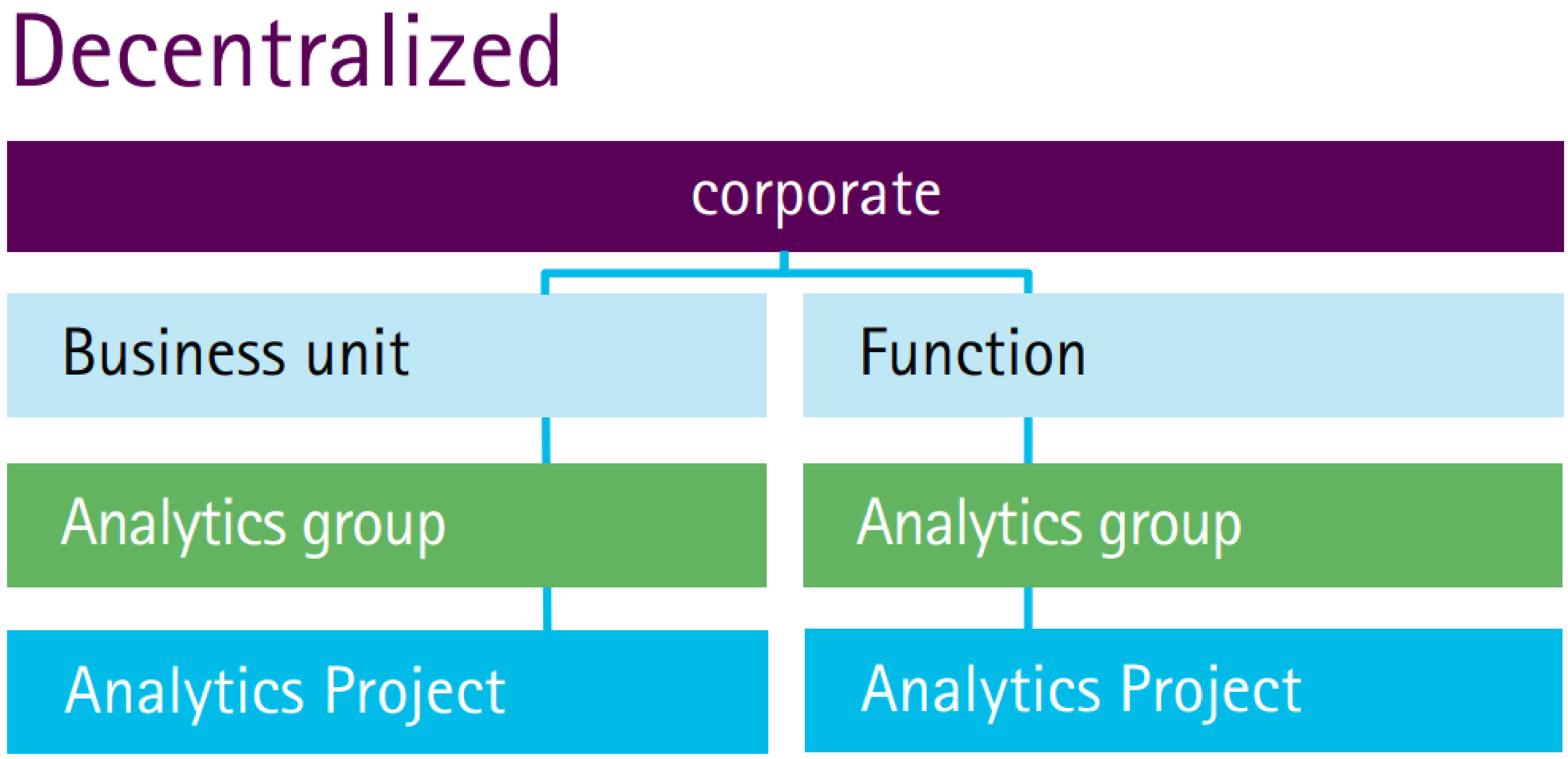
Another way to address the talent scarcity and budget limitations is to develop approachable machine learning platforms that would welcome new people from IT and enable further scaling. Even if no experienced data scientists can be hired, some organizations bypass this barrier by building relationships with educational institutions. In the US, there are about a dozen Ph.D. programs emphasizing data science and numerous boot camps with 12-month-or-so courses.

How to integrate a data science team into your company

Previously, we introduced three structure types, two of which assume that the DS team is created by combining your existing engineering effort with data science. The third specialized approach entails something more complicated, such as creating a whole new department that needs to be organized, controlled, monitored, and managed. This huge organizational shift suggests that a new group should have established roles and responsibilities – all in relation to other projects and facilities. So, how do you integrate data scientists in your company?

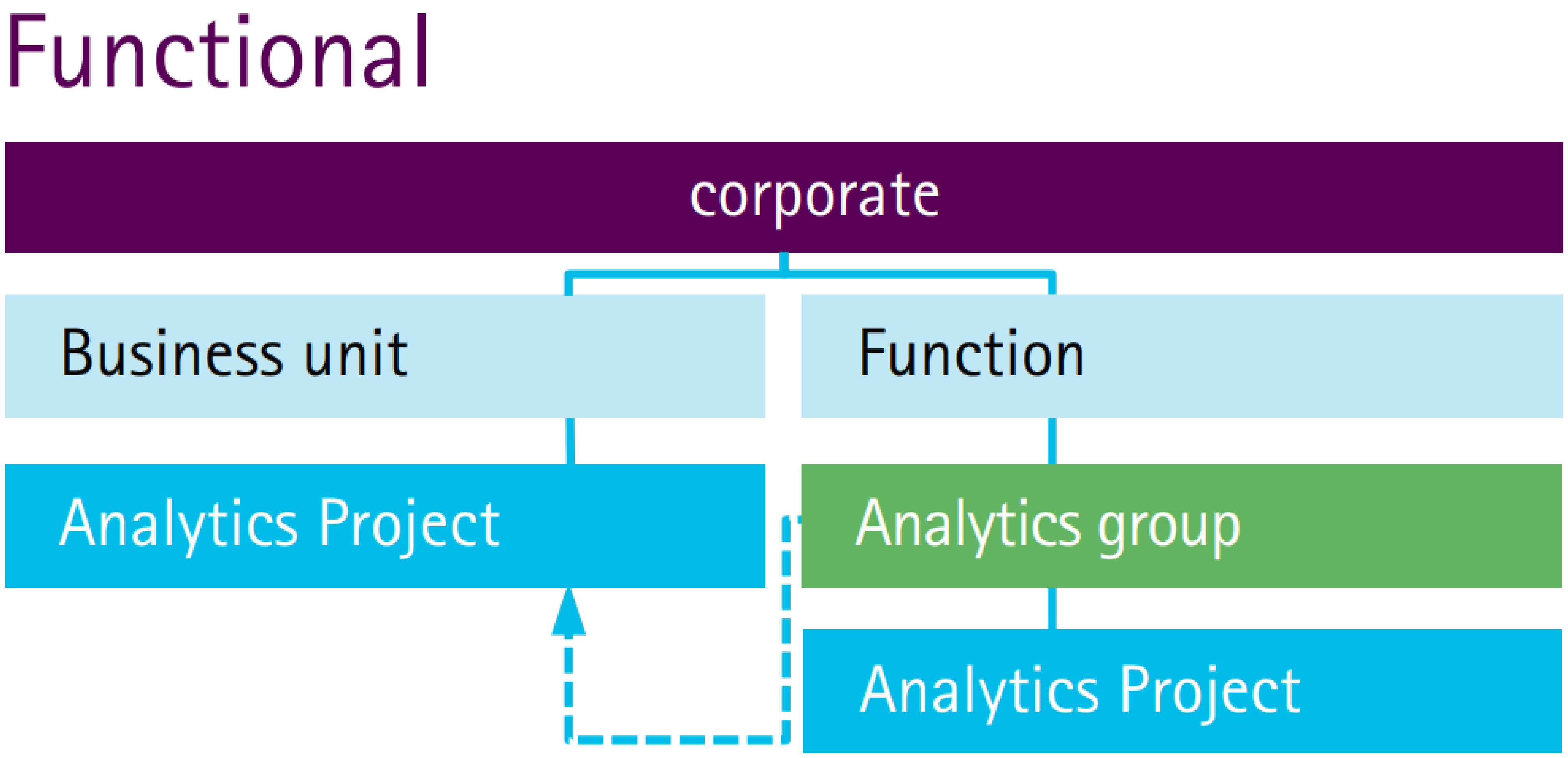
According to Accenture's classification, there are six options for organizing a data science group:

1. Decentralized. This is the least coordinated option where analytics efforts are used sporadically across the organization and resources are allocated within each group's function. This often happens in companies when data science expertise has appeared organically, which often leads to silos striving, lack of analytics standardization, and – you guessed it – decentralized reporting.



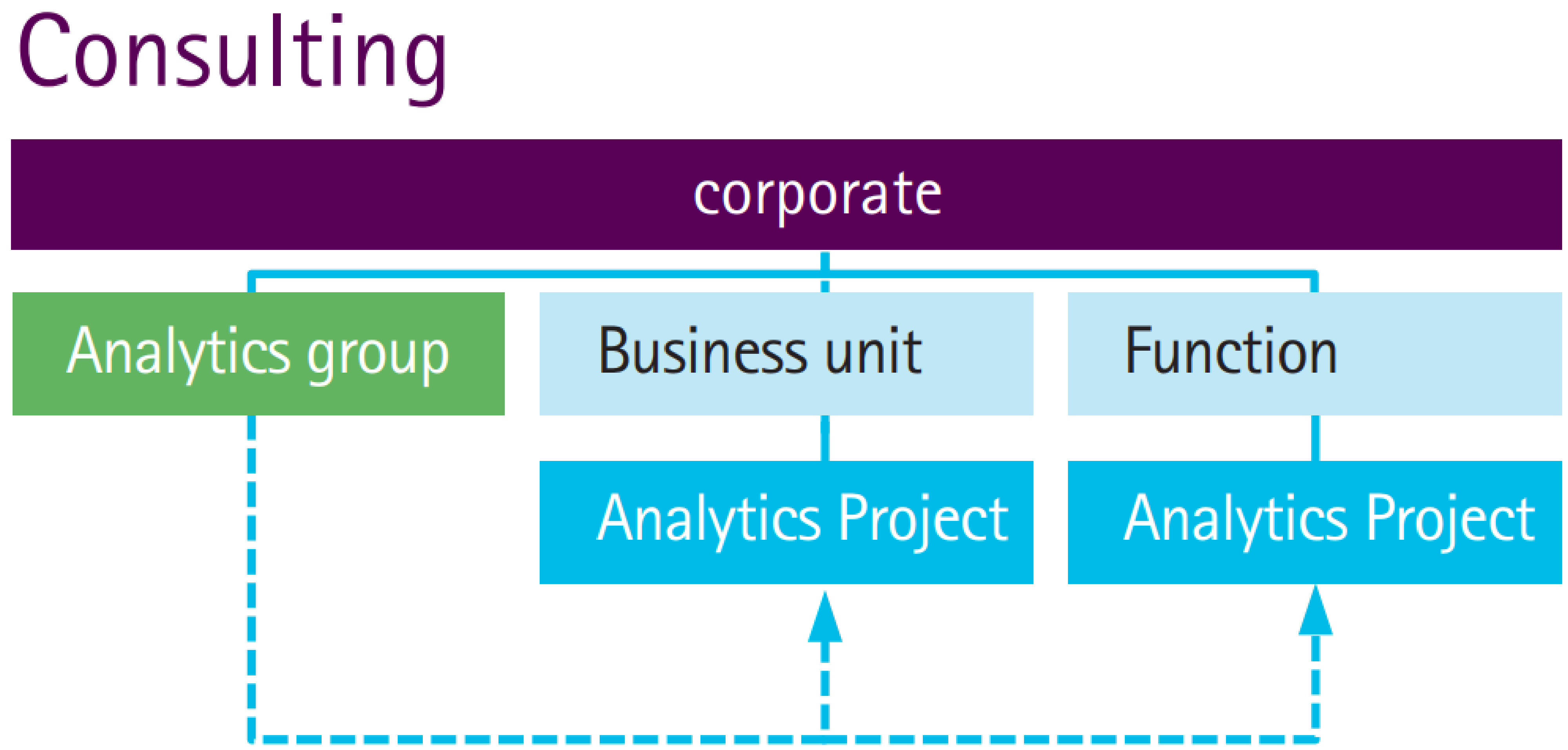
Decentralized implementation.
Image source: here and further Accenture

2. Functional. Here most analytics specialists work in one department where analytics is most relevant: it's often marketing or supply chain. This option also entails little to no coordination and expertise isn't used strategically enterprise-wide.



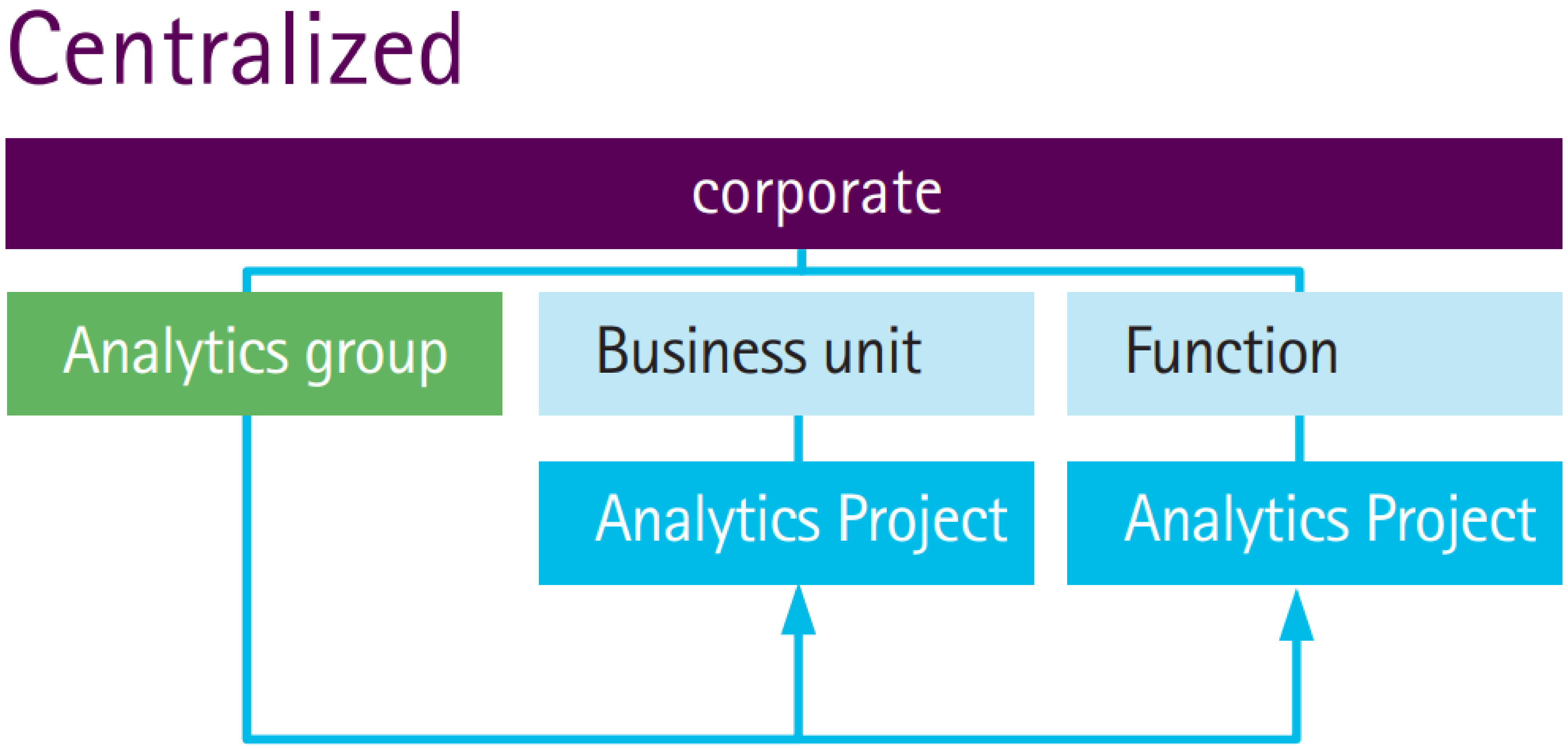
Functional implementation

3. Consulting. In this structure, analytic folks work together as one group but their role within an organization is consulting, meaning that different departments can "hire" them for specific tasks. This, of course, means that there's almost no resource allocation – either specialists are available or not.



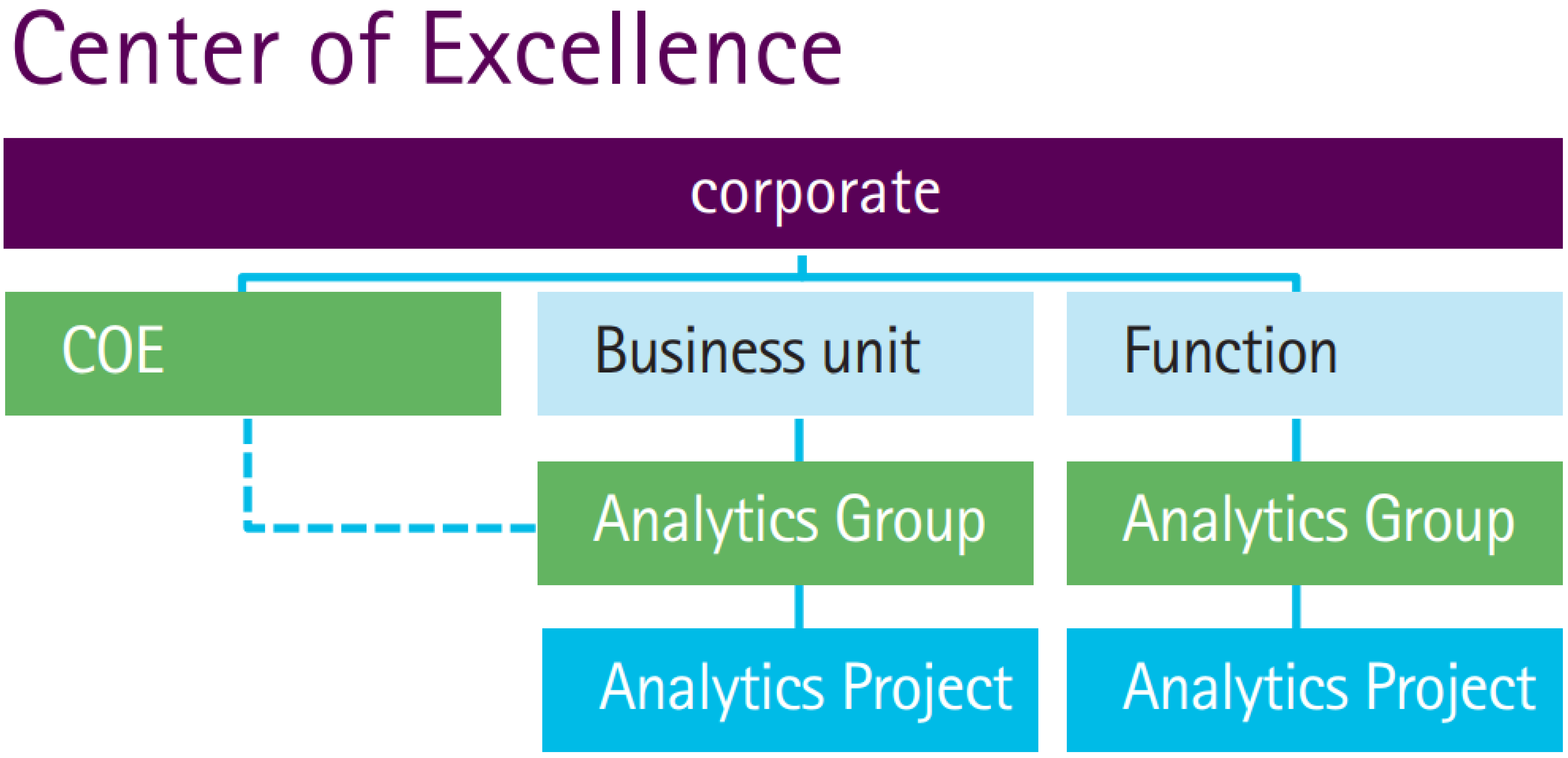
Consulting implementation

4. Centralized. This structure finally allows you to use analytics in strategic tasks – one data science team serves the whole organization in a variety of projects. Not only does it provide a DS team with long-term funding and better resource management, it also encourages career growth. The only pitfall here is the danger of transforming an analytics function into a supporting one.



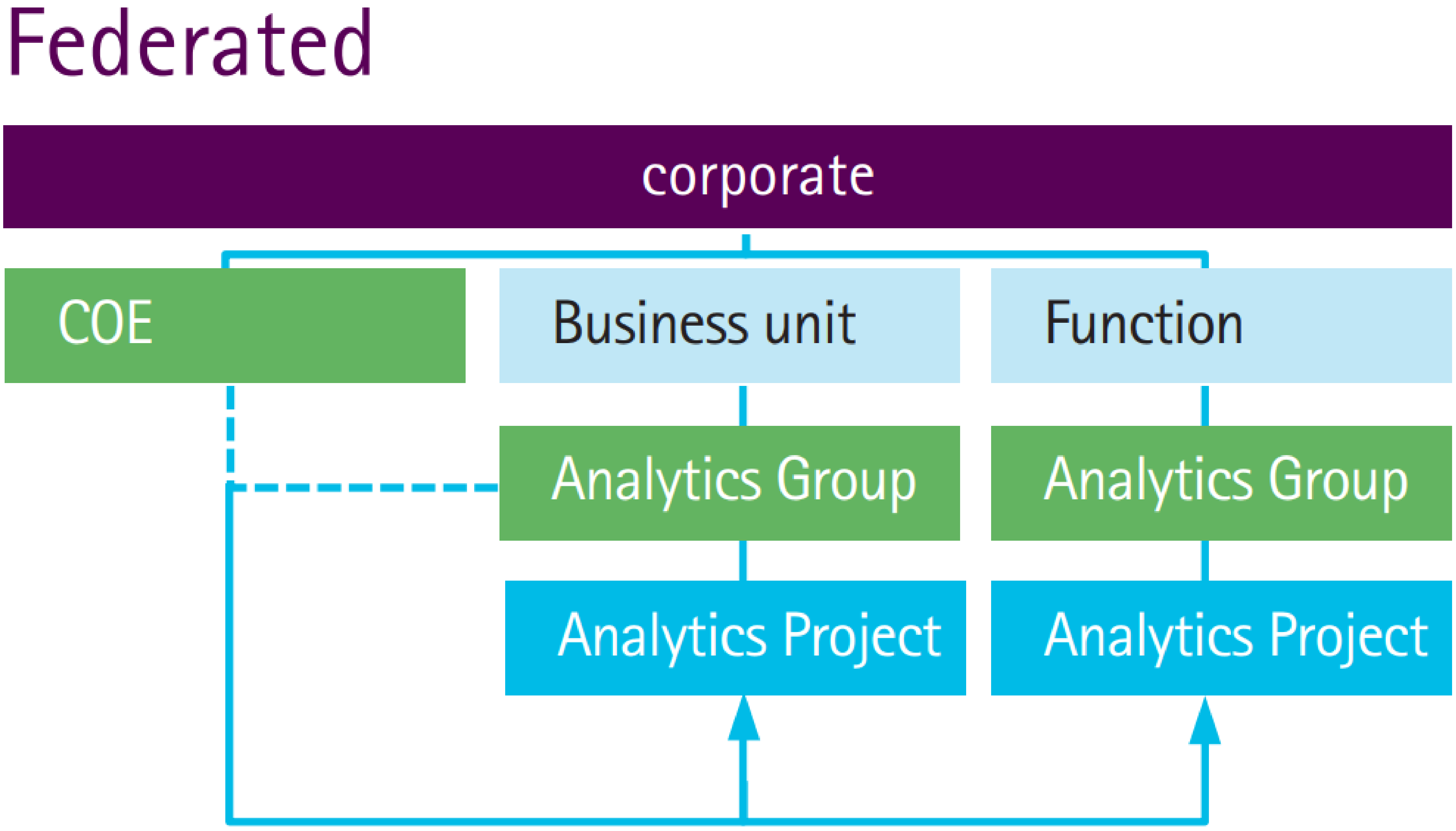
Centralized implementation

5. Center of Excellence (CoE). If you pick this option, you'll still keep the centralized approach with a single corporate center, but data scientists will be allocated to different units in the organization. This is the most balanced structure – analytics activities are highly coordinated, but experts won't be removed from business units.



Center of Excellence implementation

6. Federated. This model is relevant when there's a high demand for analytics talent across the company. Here, you employ a SWAT team of sorts – an analytics group that works from a central point and addresses complex cross-functional tasks. The rest of data scientists are distributed as in the Center of Excellence model.



Federated implementation

Remember, that your model may change and evolve depending on your business needs: While today you may be content with data scientists residing in their functional units, tomorrow a Center of Excellence can become a necessity.

More recommendations for creating a high-performance data science team

Spend less time hiring people for each title and focus on understanding what roles one individual data specialist can fulfill. For startups and smaller organizations, responsibilities don't have to be strictly clarified.

Foster cross-functional collaborations. Designers, marketers, product managers, and engineers all need to work closely with the DS team.

Practice embedding. As we mentioned above, recruiting and retaining data science talent requires some additional activities. One of them is *embedding* – placing data scientists to work in business-focused departments to make them report centrally, collaborate better, and help them feel they're part of the big picture.

Establish the team environment before hiring the team. This means that your product managers should be aware of the differences between data and software products, have adequate expectations, and work out the differences in deliverables and deadlines. PMs need to have enough technical knowledge to understand these specificities. Alternatively, you can start searching for data scientists that can fulfill this role right away.

Critical thing to be aware of

If you ask AltexSoft's data science experts what the current state of AI/ML across industries is, they will likely point out two main issues: 1. Business executives still need to be convinced that reasonable ROI of ML investments exists. 2. If they are convinced and understand the value proposition and market demand, they may lack technology skills and resources to make products a reality.

These barriers are mostly due to digital culture in organizations. Efficient data processes challenge C-level executives to embrace horizontal decision-making. Frontline managers with access to analytics have more operational freedom to make data-driven decisions, while top-level management oversees a strategy. This reduces management effort and eventually mitigates "gut-feeling-decision" risks. Basically, the cultural shift defines the end success of building a data-driven business. As McKinsey argues, setting a culture is probably the hardest part, while the rest is manageable.