Roles and Processes in Analytics Development

The rapid evolution of data analytics has been accelerated by advances in:
- large scale Internet connectivity
- data warehousing
- data analysis and mining algorithms development

We will explore data analytics processes based on these above advances. We will also look at the various roles that data analysts hold within an organization to give you a picture of what your future career in data analytics may look like.

Analytics Evolution

Data analytics spans an increasing number of industries and within these industries multiple functional areas. Data-intensive industries such as retail, financial services, healthcare, and telecommunication benefit directly from data analytics. While the increasingly large online retail industry leverages data analytics for demand forecasting and merchandising, the financial industry benefits from data analytics in credit and risk estimates, the healthcare industry in drug trials, and the telecommunication industry in product subscriptions. Within each of these industries, data analytics contributes to marketing and sales promotions, customer relationships, and distributions.

<table>
<thead>
<tr>
<th>Data-Intensive industries</th>
<th>For various functional areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Retail and Consumer Goods</td>
<td>o Marketing and Sales</td>
</tr>
<tr>
<td>o Demand forecasting</td>
<td>• Direct and digital marketing</td>
</tr>
<tr>
<td>o Merchandising</td>
<td>• Product recommendation</td>
</tr>
</tbody>
</table>
• Financial Services
  o Underwriting
  o Credit and risk
  o Fraud
  o Insurance premium and loss estimate

• Healthcare
  o Drug trial
  o New drug R&D

• Telecommunication
  o TV network
  o Product subscriptions

• Sales force optimization
  o Pricing and promotions
  • Product pricing and promotion planning

  o Customer relationship management
  • Loyalty program and lifecycle management
  • Customer acquisition and retention, cross-sell, and up-sell

  o Distribution network and supply chain

**Analytics Evolution**

Data analytics is driven by vast collections of data gathered through the Internet. It is, therefore, appropriate to align the dawn of modern data analytics, or Analytics 1.0, to that of the Internet or ARPANET in the 1950’s, soon followed by UPS’ Analytics Group. Advances in Internet accessibility and speed through router fiber optic cables, large scale storage technologies in the 2000’s created data pools that only machine learning algorithms could effectively process. This can be considered the second period, or Analytics 2.0, in the evolution of analytics corresponding to the big data era. The use of data analytics as a continuous guide to business decisions marks the era of Analytics 3.0.

DARPA was created in 1958 as the Advanced Research Projects Agency (ARPA) by President Dwight D. Eisenhower. Its purpose was to formulate and execute research and development projects to expand the frontiers of technology and science, with the aim to reach beyond immediate military requirements. The administration was responding to the Soviet launching of Sputnik 1 in 1957, and DARPA’s mission was to ensure U.S. military technology would be more sophisticated than that of the nation's potential enemies.

DARPA supported the evolution of the ARPANET (the first wide-area packet switching network), Packet Radio Network, Packet Satellite Network and, ultimately, the Internet and research in the artificial intelligence fields of speech recognition and signal processing, including parts of Shakey the robot. DARPA also funded the development of Douglas Engelbart’s NLS computer system and The Mother of All Demos; and the Aspen Movie Map, which was probably the first hypermedia system and an important precursor of virtual reality.
Analytics 1.0

Analytics 1.0, traditional analytics, or the first phase of data analytics drew from relatively small and structured data sets. The majority of the analytical activities pertained to reporting basic characteristics of the data. The mainly descriptive analytics was formed offline as “batch” processes performed on data collected the previous day. Analytics were marginal to strategy but were available to the decision-making process, still dominated by intuition.

From a technology perspective, this was the era of the enterprise data warehouse and the data mart. Data was small enough in volume to be segregated in separate locations for analysis.

This approach was successful, and many enterprise data warehouses became uncomfortably large because of the number of data sets contained in them. However, preparing an individual data set for inclusion in a warehouse was difficult, requiring a complex ETL (extract, transform, and load) process. For data analysis, most organizations used proprietary BI (Business Intelligence) and analytics “packages” that had a number of functions from which to select.
Traditional Analytics

- Data sources relatively small and structured, from internal systems
- Majority of analytical activity was descriptive analytics or reporting
- Creating analytical models was a time-consuming “batch” process
- Quantitative analysts were in “back rooms” segregated from business people and decisions
- Few organizations “competed on analytics”—analytics were marginal to strategy
- Decisions were made based on experience and intuition.

More than 90% of the analysis activity involved descriptive analytics or some form of reporting.

Analytics 2.0

- Complex, large, unstructured data sources
- New analytical and computational capabilities
- “Data Scientists” emerge
- Online firms create data-based products and services

The demand for mining the growing volume of complex, large, and unstructured data sources gave rise to Analytics 2.0. Unstructured databases (NoSQL) came to light for better management of unstructured data, and parallel servers such as Hadoop began to expedite retrieval and operations for fast-flowing data. The need for advanced analytical and computational skills prompted the emergence of “Data Scientists.” Data analytics became essential for maintaining a business competitive edge in an environment where “agile is too slow” meaning that continuous development is preferred over the typical two-week agile cycle, and where the role of a consultant generating reports is superseded by unbiased, real-time computer analysis.

Key Developments

- Fast flow of data necessitated rapid storage and processing
- Parallel servers running Hadoop for fast batch data processing
- Unstructured data required “NoSQL” databases
- Data stored and analyzed in public or private cloud computing environments
- “In-memory” analytics and “in-database” analytics employed
- Machine learning methods meant the overall speed of analysis was much faster (from days to minutes)
- Visual analytics often crowded out predictive and prescriptive techniques
- “Agile is too slow”
“Being a consultant is the dead zone”
Information (and hardware and software) wants to be free
Share your big data tools with the community

Analytics 3.0

There is considerable evidence that large organizations are entering the Analytics 3.0 era. Such is the era where analytics is integral to running a business and an essential aspect of strategic planning. Trends and analyses instantly obtained for time intervals of interests provide rapid and agile insights. It’s an environment that combines the best of 1.0 and 2.0—a blend of big data and traditional analytics that yields insights and offerings with speed and impact.

Characteristics

- Analytics integral to running the business; strategic asset
- Rapid and agile insight delivery
- Analytical tools available at point of decision
- Cultural evolution embeds analytics into decision and operational processes
- All businesses can create data based products and services

"...virtually any type of firm in any industry can participate in the data economy."

Analytics Applied

Learning Objectives
- Explain the technical elements and steps associated with analytics practices and processes

The ladder of sophistication in data analytics can be viewed as the sequence of exploratory analysis, predictive analysis, prescriptive analysis, and optimization. Exploratory analysis pertains to the extraction of summary descriptive analytics; predictive analytics provides the capability to forecast based off of past information and trends; prescriptive analysis turns information from the previous two steps to provide actionable information, and optimization considers business and data derived constraints to suggest ‘best’ courses of action.
Analytics Applied at Various Sophistication Levels

Versions of “Translation” of “Analytics”

Business intelligence combines a broad set of data analysis applications, including ad hoc analysis and querying, enterprise reporting, online analytical processing (OLAP), mobile BI, real-time BI, operational BI, cloud and software as a service BI (SaaS BI), open source BI, collaborative BI, and location intelligence. BI technology also includes data visualization software for designing charts and other infographics, as well as tools for building BI dashboards and performance scorecards that display visualized data on business metrics and key performance indicators in an easy-to-grasp way. BI applications can be bought separately from different vendors or as part of a unified BI platform from a single vendor.

<table>
<thead>
<tr>
<th>Functions</th>
<th>Data modeling and reporting</th>
<th>Business intelligence and visualization</th>
<th>General data analysis</th>
<th>Advanced analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>o ETL (Extraction, Transformation, and Loading) process</td>
<td>o Data reporting (KPI's, metrics)</td>
<td>o Data reporting (KPI's, metrics)</td>
<td>Example: o Insurance auditing</td>
<td>o Complex, large-scale analysis involving a good amount of data and Predictive Analytics / Data</td>
</tr>
<tr>
<td>o Data storage (data warehouse, data lake)</td>
<td>o Dashboard</td>
<td>o OLAP (cube, slice, and dice)</td>
<td>o Product quality control</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>o Ad hoc data analysis</td>
<td>o Network traffic analysis</td>
<td></td>
</tr>
</tbody>
</table>

Example:
- Insurance auditing
- Product quality control
- Network traffic analysis

Example:
- Complex, large-scale analysis involving a good amount of data and Predictive Analytics / Data
Data modeling (schema design)
- Automatic monitoring
- Operational, real-time BI
- Little or no machine learning applied
- Mining / Machine Learning
  - Credit risk scoring
  - Marketing mix modeling

Tools
- Oracle SQL | MS SQL
- MySQL Postgre SQL
- IBM Cognos | Oracle OBIEE
- More tool options
- Excel | Various tools
- R / Python / SAS / SPSS
- Various Tools

Roles
- IT department
- DB analyst / Data Engineer
- IT department
- Data Analyst / Business Analyst
- Data Analyst / Financial Analyst
- Statistician / Modeler / Data Scientist

Fundamental Elements in Analytics Development

- **Data**: A data warehouse only stores data that has been modeled/structured, while a data lake is no respecter of data. It stores it all—structured, semi-structured, and unstructured.

- **Processing**: Before we can load data into a data warehouse, we first need to give it some shape and structure—i.e., we need to model it. That’s called schema-on-write. With a data lake, you just load in the raw data, as-is, and then when you’re ready to use the data, that’s when you give it shape and structure. That’s called schema-on-read. Two very different approaches.

- **Storage**: One of the primary features of big data technologies like Hadoop is that the cost of storing data is relatively low as compared to the data warehouse. There are two key reasons for this. First, Hadoop is open source software, so the licensing and community support is free. And second, Hadoop is designed to be installed on low-cost commodity hardware.

- **Agility**: A data warehouse is a highly-structured repository, by definition. It’s not technically hard to change the structure, but it can be very time-consuming given all the business processes that are tied to it. A data lake, on the other hand, lacks the structure of a data warehouse—which gives developers and data scientists the ability to easily configure and reconfigure their models, queries, and apps on-the-fly.

- **Security**: Data warehouse technologies have been around for decades, while big data technologies (the underpinnings of a data lake) are relatively new. Thus, the ability to secure data in a data warehouse is much more mature than securing data in a data lake. It should be noted, however, that there’s a
significant effort being placed on security right now in the big data industry. It’s not a question of if, but when.

- **Users:** For a long time, the rallying cry has been BI and analytics for everyone! We’ve built the data warehouse and invited “everyone” to come, but have they come? On average, 20-25% of them have. Is it the same cry for the data lake? Will we build the data lake and invite everyone to come? Not if you’re smart. Trust me, a data lake, at this point in its maturity, is best suited for the data scientists.

### Roles and Processes

Analytics development roles can generally fall into one of the four groups: business analyst, modeler/data scientist, IT system manager, and business manager. The business analysis defines the business requirements, performs some data exploration then creates visualizations and reports to support and evaluate the requirements. The data modeler/scientist seeks to create predictive models and mine the data for unforeseen relationships. The IT system manager is responsible for data preparation, model deployment, and solution implementation. In other words, the IT system manager manages the entry of data, the internal application of models to the data, and the customer-facing solutions.

### Roles in Data Analytics

<table>
<thead>
<tr>
<th>Business manager</th>
<th>Business analyst</th>
<th>IT system manager</th>
<th>Modeler/Data Scientist</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Domain expert</td>
<td>• Business requirement</td>
<td>• Data preparation</td>
<td>• Data mining</td>
</tr>
<tr>
<td>• Manage processes</td>
<td>• Data exploration</td>
<td>• Model deployment</td>
<td>• Predictive modeling</td>
</tr>
<tr>
<td>• Monitor performance</td>
<td>• Visualization and report</td>
<td>• Solution implementation</td>
<td>• Analytical solutions</td>
</tr>
</tbody>
</table>
Processes in Data Analytics

The processes in analytics development can be described as a cyclic pattern which typically begins with the identification of a problem or potential benefit from a data set. Once the data set is identified, it is extracted and prepared, or "cleaned up," for exploration. During the exploration process, the analysts develop seed ideas that will subsequently guide the modeling and solution-building tasks. Once the models are validated, the model is prepared for deployment on new (or live) data. This process will often repeat itself.

Steps in an Analytics Development Process

<table>
<thead>
<tr>
<th>Problem definition</th>
<th>Define the analysis goal, scope, end deliverables, and potential approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data requirement</td>
<td>Identify the required data sources, data fields, and data coverage (e.g., temporal and geographical coverage)</td>
</tr>
<tr>
<td>Data collection</td>
<td>Extract, gather data from one or multiple sources and transform them into a consumable format</td>
</tr>
<tr>
<td>Data cleansing</td>
<td>Identify and fix the data errors, outliers, any unwanted data elements</td>
</tr>
<tr>
<td>Data processing</td>
<td>Transform the data format or data variables to meet the needs of downstream analysis, modeling, or development activities</td>
</tr>
<tr>
<td>Model building</td>
<td>Develop algorithms and build computer models based on the goal of analysis</td>
</tr>
</tbody>
</table>
**Model deployment**
Generate the results and insights from the models and implement the models in a systematic manner

**Communication**
Communicate the results and insights to the business users and make recommendations on subsequent actions

### Typical Analytics Project Workflow

A typical analytics project workflow begins with the identification of a problem. The data preparation steps to follow include the data requirements and access, the data extraction, transformation, and loading (ETL), as well as data sampling (selecting a representative subset). During the model development phase, the models are built and validated. The model deployment concludes the workflow. It should be noted that significant effort on the workflow is absorbed in the data preparation phase.

![Diagram of Typical Analytics Project Workflow]

### Summary | Roles and Processes in Analytics Development

Data analytics spans an increasing number of industries and within these industries are multiple functional areas. Data-intensive industries such as retail, financial services, healthcare, and telecommunication benefit directly from data analytics. Within each of these industries, data analytics contributes to marketing and sales, promotions, customer relationships, and distributions. Each industry sector using data analytics has its own roles, with even more specificity within organizations, that all contribute in unique ways to the collection and analysis of data.

The process of data analytics is applied in a variety of ways and at different levels of sophistication. The application of predictive analytics, providing the capability to forecast based on past information and trends, will be examined next.