XONA PARTNERS

Data Science

A Practical Perspective

Dr. Riad Hartani, Dr. James Shanahan,
Dr. Anurag Maunder,
Dr. Alex Popescul (Xona Partners)
Dr. Gopal Dommety (N42, Inc.),
Jagadish Channagiri (Moogilu, Inc.)

February 2014

1 Preamble

As businesses evolve to leverage the huge amounts of data assembled - mining and learning through such data as well as optimizing communication between those producing it and those using it brings about the golden age of applied Data Sciences. This development touches upon the evolution of the compute platform over which data sets are resident and addresses aspects as varied as data management, reliability & availability, performance & scalability, real-time analytics, data API and governance. Data Science addresses all these complexities by organizing data, building data infrastructure, and extract business logic semantics

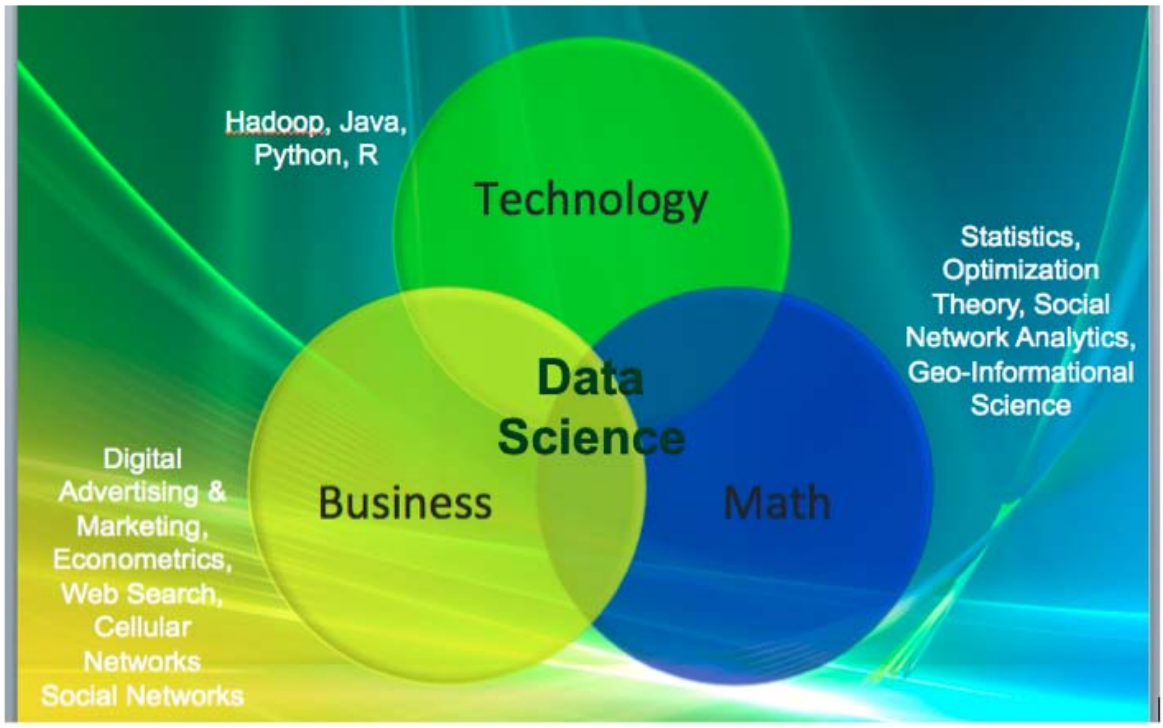
Unlike many other Data Science articles that cover the basics and technological underpinnings, this paper takes a practitioner's viewpoint. It identifies the use cases for Big Data Science, its engineering components, and Big Data Science integration with business processes and systems. In doing so, it respects the large investments in data warehouse and business intelligence and shows both evolutionary and revolutionary— as well as hybrid— ways of moving forward to the brave new world of Big Data Science.

2 Big Data Science – Context & Synopsys

Why is Big Data different from any other data that we have dealt with in the past? There are "four V's" that characterize this data: volume, velocity, variety, and veracity. Throughout this paper, we will revisit some of the considerations that related to the "four V's", in terms of gathering data, storing it and accessing it dynamically, organizing it and leveraging it. This will be mostly done via illustrative practical use cases.

Through various use cases conducted for organization with strategic data transformation projects, we generally run into five types of Big Data: web and social media, machine-to-machine (M2M), big transaction data, biometrics, and human generated. Here are some examples of Big Data that we will use in this report: social media; text cell phone locations; channel click information from set-top box web browsing and search; product manuals communication network events; call detail records (CDRs); radio frequency identification (RFID) tags; and application interactions. Addressing and leveraging such data sets is at the core of the Data Sciences realm.

Data Science, a relatively new discipline, is a combination of technology, business and mathematics that increasingly impacts every facet of daily life. The combination of traditional disciplines of data extraction (ETL), data intelligence, data analytics, data modeling, data warehousing, and reporting along with statistics and predictive analytics can be referred to as Big Data as illustrated in the diagram below.



A natural way of visualizing the various components of the Data Sciences hierarchy is shown below, taking it from data extraction at the bottom all the way to applications to specific verticals at the top.

Solutions (Healthcare, Internet, Security, Mobile and more)

Lifetime Modeling (action-based)

Realtime (Scoring, AB Testing, DOE, Event logging)

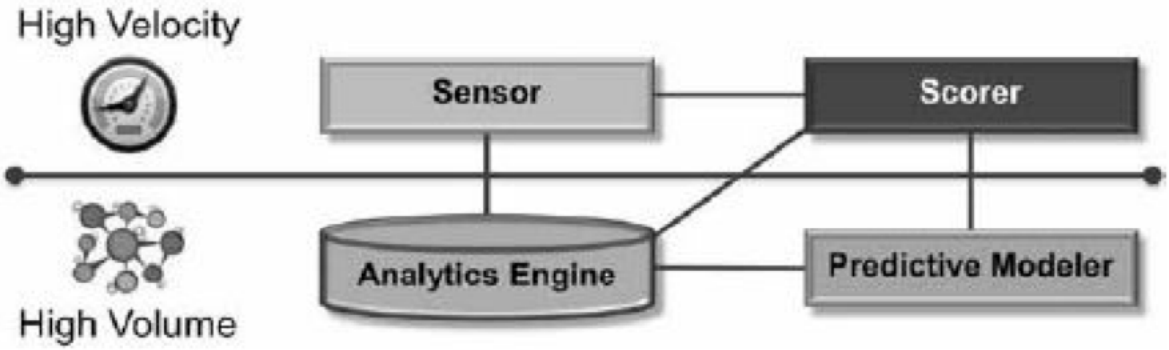
Visualization +UI (Dashboards, Admin Tools, Reporting, ad creation and targeting)

Analytics (Profiles, Machine Learning, Artificial Intelligence)

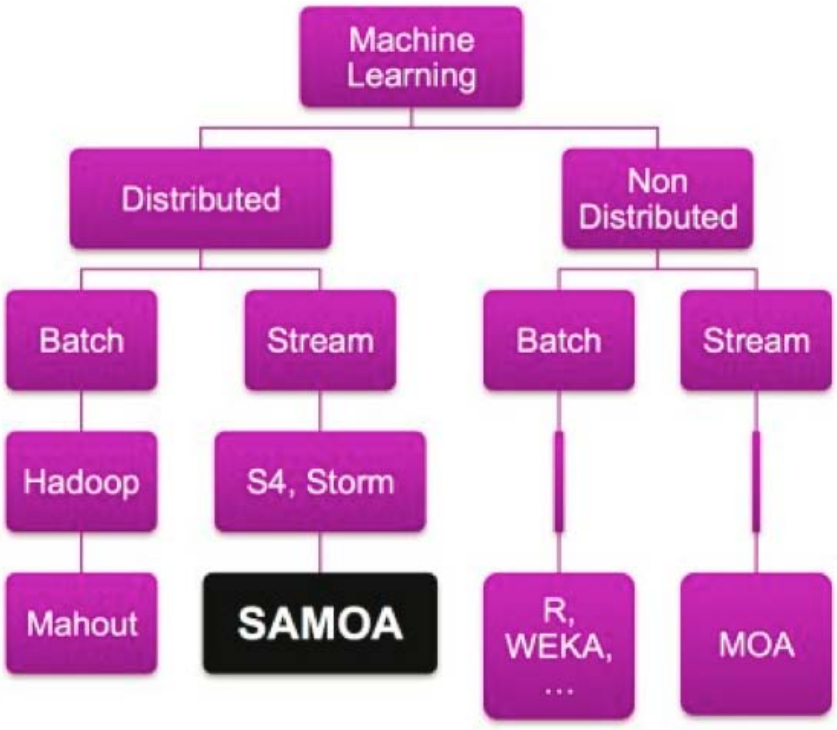
Data Store (HSFS/RDBMS/Real-time Stores)

Large Scale Data Capture

Another way to illustrate the flow of logic within a Data Science framework is illustrated below, which would include the closed loop feedback formed between data acquisition and learning through data models.



Big Data has many moving parts and needs a team that can synthesize information to build a solution that helps solve a customer problem. In the diagram below, machine learning solution SAMOA is described.



SAMOA (Scalable Advanced Massive Online Analysis) is an example of framework (we have used internally in some of the use cases described below) for mining big data streams. As in most of the big data ecosystem, it is written in Java. It features a pluggable architecture that allows it to run on several distributed stream-processing engines such as Storm and S4. SAMOA includes distributed algorithms for the most common machine learning tasks such as classification and clustering. For a simple analogy, you can think of SAMOA as Mahout for streaming.

Before SAMOA can be used, the data has to be structured and made available on Hadoop environment. The data can be organized and streamed to Hadoop environment using traditional ETL techniques – database streaming and file system streaming. And the data infrastructure consists of servers and storage arrays. Hadoop runs on the server and provides reliability by data redundancy and performance is provided through data distribution.

SAMOA as a distributed Machine Learning framework works on the data set imported into Hadoop environment. SAMOA uses the data set for training and execution. For example, SAMOA can be trained in near real-time to detect SPAM. The algorithm can be applied in real-time to new data streams to detect SPAM. SAMOA output can be streamed to higher layers for visualization and reporting. The entire Big Data flow and Data Management is described in Section 3.

3 The Big Data Management Angle

3.1 The Challenges

A preamble to leverage any big data framework is to be able to manage the various flows of data throughout their lifetimes. This includes the ability to identify, capture, and manage data to provide actionable and trusted insights that improve strategic and operational decision making, resulting in incremental revenues and better customer experience. As such, the desired goal is to create a solid foundation architecture that is able to provide these optimal functional capabilities and a platform to overlay Data Science applications.

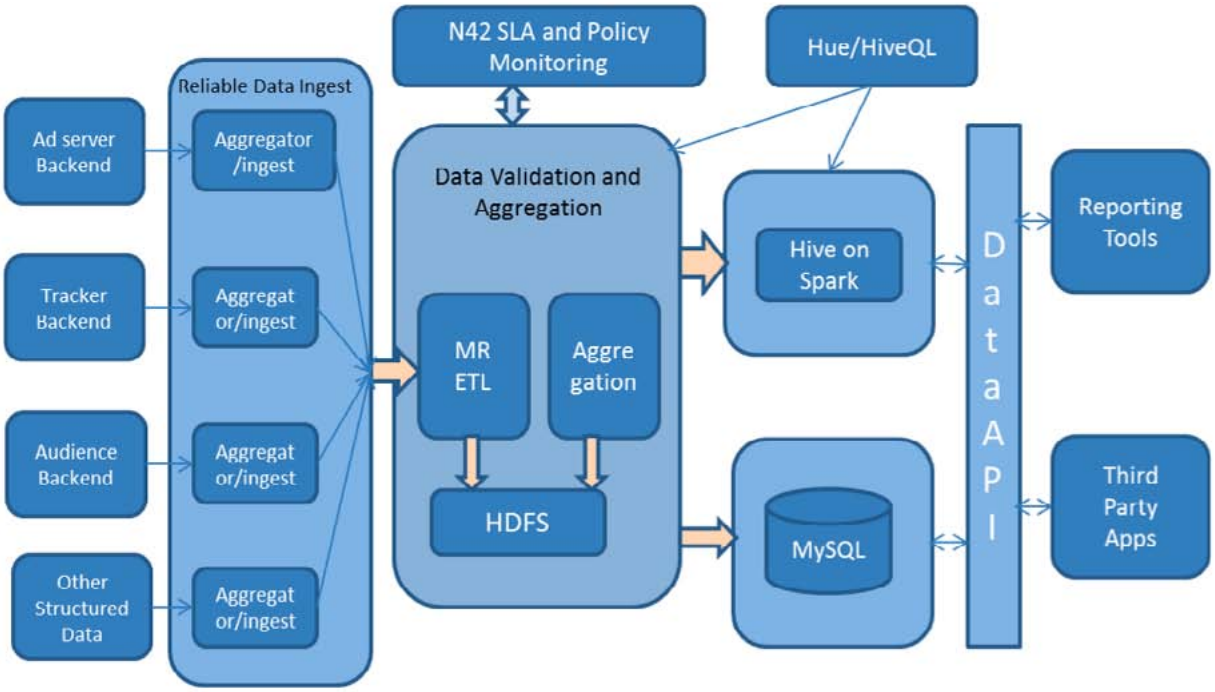
The current challenges of the existing data management platforms mostly affect the operation team's ability to provide reliable control over critical aspects such as the scalability of the data platforms to perform effectively as the business data volume and messaging grows. As such, the desired goal is to create a solid foundation architecture that is able to provide these optimal functional capabilities and a platform to overlay additional applications such as business intelligence and Data Science as a service capability.

Today, the existing information infrastructure and analytics processes suffer from challenges we have observed in most typical large-scale data projects, some of which are listed below. Data transformation projects' main focus is on providing solutions to these challenges:

- Fast changing data sets in terms of structure and content
- Hard to meet stringent SLAs for data feeds handling, access, aggregation and queries
- Data architecture complexity in terms of naming, change management, and analytics
- Lack of comprehensive policy-based management for retention, replication, archival, and compression
- Difficulty in providing real time granularity for real time reports generation
- Complexity in aggregating data sources from multiple data centers as dictated by real time processing constraints

3.2 Novel Data Management Framework

Data architectures are fast evolving with the goal of addressing the various challenges described above. Here below is a proposed architecture:



We describe some of the major trends and a set of possible software implementation of such functionalities over the next few years. These are provided as examples, noting that various other implementation techniques are available.

Data Storage and Warehousing: Various implementation frameworks have been implemented. Hbase, as an illustrative example, is very efficient in fast time range scanning; time range queries, and data drilldown, in the face of read-only data with low throughput write data. Additionally, HBase supports quick snapshots and is an ideal data-warehousing platform. Data cubes stored in HBase allow cube operations such as pivoting, and drilldown via HBase. HBase is a good data warehousing option in terms of cost/performance for report generation. In a similar way, Hive on Spark allows for in memory queries for analytics that provide near real time analysis of data. This component will address SLA requirements of the reporting solution without having to implement the existing reports but with added performance. Additionally, this provides better data import/ export than, for example, MongoDB noSQL solution with better performance for lower cost.

Optimization of Data Architecture Availability and Reliability: Here again, Hadoop 2.0 and upcoming iterations of Hadoop, as an example, will form the basis of the availability and reliability architecture. It supports distributed Jobtracker and high availability to Datanode. This avoids single point of failure for the Hadoop deployment. HDFS replication itself lends to high availability of data on file system. Zookeeper should be implemented with multiple nodes for high availability of cluster. Data policies for archiving and snapshots through HBase will provide reliability and disaster recovery options for the cluster. Data ingest is one of the critical first steps in achieving data consistency for analysis. Flume allows predictable and efficient data ingestion

into HDFS file system providing visibility into failures and improving performance of data ingestion. Missing data can be detected with custom plugins to Flume pipeline. Depending on the requirements, it is possible to use Kafka in the pipeline for reliable delivery of data for preventing data loss. In a similar way, Oozie provides event based workflow mechanism for launching jobs in the event of data ingest into HDFS or HCatalog. Additionally, Oozie provides an easy way of specifying job workflow including Pig and Hive jobs allowing SLA specification for workflow. This implementation will allow better quality of data for reliable reports and better performance on scheduled reports as well as ad-hoc queries.

Data Management Performance and Scalability: Cloud based deployments will form the basis of the scalability models for the IT and backend architectures. Here again, and as an example, performance and scalability improvements are achieved using Hive on Spark. Linear scalability and performance with scale can be achieved by using the Hadoop 2.0 architecture as defined in the next section. This cluster is designed to be single cluster to support data needs that can consolidate all or some, of its data centers. Processes and policies in place for data lifecycle management for archiving, retention, compression and replication will allow for efficient data management with low overhead costs.

Data API and Exposure to 3rd Party Applications: Data API is a virtualization layer that hides underlying platform details and provides REST or JDBC interfaces for external interaction. There will likely be an evolution towards the integration of on and off premise Data API solution providers, natively or as SaaS model. Some solutions simplify data import export to noSQL databases. These solutions can be integrated to provide a consistent data view to external actors. Application development using the data interfaces that are decoupled with data storage structure will lead to lower cost of maintenance and better integration with partners.

Real-Time Analytics: Real-time data processing has to accommodate high velocity data stream and process data in near real time for alerts and analysis. Real time processing systems, using frameworks such as storm and Kafka will allow for horizontal scaling, large-scale events processing, reliable data management and dynamic events handling. Storm supports high throughput event processing and achieves reliability using Kafka for incoming data. Processed events can generate events that can be acted upon for real time processing by additional jobs. The processed data is persisted using HBase for efficient storage and can be combined with historical data in the cluster for generating reports at regular intervals. This real-time processing infrastructure will support mobile reports that are expected to be generated in near real-time, which in itself is a great value add as far as intrinsic business value.

The implementation of novel data management models, at different levels of the data flow hierarchy, with some examples highlighted above, leads to significant improvements in the overall business value. In some of the use cases analyzed, and taking into account various KPIs to measure such improvement, as dictated by the specific vertical and business problems under consideration, we did observe a direct correlation of data management processes put in place and the resulting efficiency optimization.

4 The Intelligent Analytics Component

At the core of the overall Data Science framework sits the intelligent data analysis component, which leverages an extensive set of machine learning and data modeling techniques.

As a team, we have first contributed to the field of artificial intelligence and overall data and semantics modeling since the early 90s with early work in the area of machine learning, multivalued logic and neural networks. We then followed the evolution of these data analysis and knowledge discovery techniques over the years, as algorithms became more elaborate, computing models became more efficient, and live data is generated and collected at increasingly higher rates, often for completely novel applications.

Very recently, our focus has been on analyzing applications of recent techniques such as deep learning and the various additions to random forest and gradient boosted decision trees to practical industry problems. The intelligent data analysis and logic based reasoning techniques have made the big leap of being a research area for a select few applications, to a set of tools, accessible in various shapes and forms to various industry verticals, and optimized to resolve some of their more challenging problems.

We are now witnessing the emergence of enhanced (and some cases new) set of machine learning and data mining algorithms, specifically focused on clustering and predictive modeling in high dimensional spaces based on imprecise, uncertain and incomplete information, efficient statistical data summarization and features extraction algorithms as well as large-scale real-time data stream management. These tools will be at the core of the processing engines being commercialized or running in open source environment, and will aim, when applied to specific industry problems, at optimizing the existing business logic and augment it with new functionalities over time.

5 Applications & Case Studies

We have been working on various case studies in different verticals, including mobile Internet, financial applications, healthcare, and online advertising. The focus has been on leveraging large-scale network datasets to apply machine learning-based analysis of various network activities with the aim to provide significant improvement over standard techniques used today.

5.1 Generic Use Cases Characteristics

Addressing these various use cases, we opportunistically leverage the fact that, when it comes to data management, a few concurrent trends are converging. A brief snapshot is presented here.

First is the maturity of data management models

We are witnessing the fast adoption of novel architectures to store and access large data sets (Hadoop, MapReduce, HDFS, Yarn, etc. – commonly known as Big Data models), as well the commercial availability of various cloud deployment architectures (OpenStack, vCloud, Cloudstack, AWS, etc.). This is removing significant logistical obstacles to embracing management of large data structures. The move is likely to be even more significant moving forward, given the immense number of contributions of the open source community in this area. The key here is convergence onto universally adopted platforms versus what was before seen as a proliferation of diverse platforms.

Second is the evolution of Data Sciences

This applies to the large set of data analysis models in a broad sense, and specifically machine learning and mining algorithms that are more accurate and computationally tractable, leveraging

distributed cloud-based computing models. Current developments in Deep Learning, for example, illustrate well how an older field of neural networks achieved breakthroughs in accuracy when its algorithm improvements were fueled by much increased computational power. Taking advantage of the introduction of new computing models, such as algorithms parallelization, GPUs and alike, then porting that to distributed cloud compute models, not only the existing algorithms have been optimized to run better and faster, but a number of additions and optimization have been developed and run in a computationally tractable way.

Third is Data Availability

Leveraging compute and storage architectures that are increasingly scalable to selectively and dynamically process large volumes of data, relying on various models of data capture, via sensors, devices, and management modules. Larger data sets influence algorithm choices by easing the risks of over-fitting, which leads to better generalizable insights. The sheer size of data available is likely to increase, either as front-end data in real time or backend data stored as historical patterns. In the financial world specifically, data collection architectures have evolved in a way that allows for data to be captured fast enough for deeper analysis, and software-based data management architectures in a way that data can be queried, received and presented to relevant data processing models.

5.2 Mobile Network Optimization

3G and 4G networks are built over flat IP packet-based networks. With the flexibility and scalability of IP-based networks and services come the requirements for more stringent traffic and resource management mechanisms and underlying challenges unseen in previous circuit-based switching technologies The new architecture introduces various network elements in order to tackle such challenges. This would include data-path processing models such as Deep Packet Inspection devices, used for marking and rate limiting traffic, to data compression/rating devices used for video optimization, to topology and state aware control-plane devices for resources load-balancing engines, among others.

In order to optimize customer user experiences measured by Quality of Experience (QoE: defined along various KPI metrics as perceived by the user), 3G/4G networks require the introduction of more sophisticated predictive, preventive and/or corrective resource management models in the network. This is specifically where we have introduced novel data-processing models, leveraging machine-learning algorithms, and demonstrated their value. As such, a real-world traffic control scenario is developed addressing a very specific problem that is causing major challenges in mobile networks today. The problem is formulated as follows: How to maximize the aggregated-user QoE utility function over time based on observation of real-time and batch historical network-level data measurements, thereby enact semi-real-time traffic control mechanisms at specific network enforcement points, either directly through dynamic provisioning or via a policy proxy function, such as a PCRF.

5.3 Fraud, Revenue Assurances and mobile payments in Mobile networks:

As mobile operators put billing and revenue settlements in place, various inconsistencies, due to multiple reasons spanning from network configuration errors to various fraud cases, are noticed.

The process to cluster these cases into specific clusters, and categorize them based on the reasons why they happened is based on various identification models, most of them rule based. It is improvement to such models, leveraging Data Science techniques that we aim at addressing. The same would apply to model required to identify various mobile customer categories based on their historical profiles and induce through that, the billing and control models to put in place when activated into the network.

Various extensions of this case study to mobile payment scenarios, where mobile payment data, along with the underlying mobile payment logic (via mobile operator payment network elements, or third party proxy payments via financial institutions or alternatives) will potentially be considered based on the initial analysis of mobile operators' feedback as far as addressing these new challenges.

5.4 Mobile advertising optimization models

As online advertising business continues its growth path (providing the core of the revenue streams of leading Internet players such as Google, Microsoft, Apple, Facebook, etc.), and mobile operators aim at sharing some of the revenue pie, along diverse business models, and doing so, require the introduction of new models to understand mobile users data and how to leverage them so they can be part of the overall mobile advertising value chain. Overall, the key problem to be tackled is to increase the revenue per user for advertisers, with mobile operators aiming at capturing a piece of such revenue, either through a revenue sharing model or through a direct revenue generation.

It's the revenue per user that would form the main metric of optimization, when comparing these numbers for various OTTs (Over the Top). Having mobile operators increase this revenue through various schemes is what would form the basis for new services, or new business models, with a direct consequence on the products and solutions strategies of the mobile service providers. This as well as optimizing bids of DSPs through RTB on Ad exchanges, with the complementary goal of optimizing CPMs (Cost per Mille), CPAs (Cost per Action) and CPCs (Cost per Click). Based on large data sets available, one will aim at extracting the optimal underlying semantics, that would be leveraged in increasing the accuracy of mobile advertising targeting, reduce fraud in mobile advertisements insertion and develop a model where each layer of the mobile advertising value chain, would optimize its returns.

5.5 Healthcare Optimization Models

"Prevention is better than cure" is the famous quote by 15th century philosopher Desiderius Eramus and is absolutely true in the modern healthcare. To categorize and capture health issues at a very early stage in one's life is critical to bring down healthcare costs. The US government funds healthcare programs that are administered by States at every school to help underprivileged children with health issues. This helps individual States to track health issues at a very early stage and prevent this cascading into a much more serious problem at later stages. Here is where the Big Data will be of enormous help. The data of current student's health is in silos of individual provider. The data captured includes: the nature of disease, treatment, location, and costs. The data can be ingested and can be used to derive foundational analytics. The models can be set up to correlate the effects of early stage preventions; model can predict the states/counties at risk.

The models can predict risk profiles based on ethnicity, age, sex, city/rural dwellers, and socio-economic factors. The Big Data Science service can help reduce healthcare costs. Another major issue of healthcare is "hotspots" – some groups consume most of healthcare costs. With Big Data Science - analyzing medical records, one can detect current hotspots and also predict the future hotspots. This can help reduce healthcare costs.

5.6 Mobile network security optimization models

Event-based models largely drive current approaches to information security and information assurance in the mobile telecom sector. These models, which largely rely on human interaction and reactive experience, attempt to correlate events across heterogeneous platforms, and are notoriously inappropriate when faced with rapidly changing threat vectors, variant information asset values and increasingly rapid technology cycles. As individual platform types can operate in signature-based, anomaly-based or heuristic models, very little end-to-end perspective is available to organizations that rely on network platforms as core operational assets, and risks are frequently only recognized in hindsight. As the information security and information assurance exposure scales up, Operators are increasingly searching for a more dynamical and predictive alternative to mitigate these critical risks. Such a behavioral model operates on 3 primary facets: (1) Subscriber-facing attack surfaces & behavior, (2) Internet-facing attack surfaces & behavior and (3) Ecosystem- & partner-facing attack surfaces & behaviors. An IS&IA model predicated on causality rather than pattern matching is a compelling alternative to the shortfall of the current methodologies, and further opens the door to revenue-generating services to downstream subscribers and customer organizations.

6 Conclusions

This paper addressed the inter-related fields of Data Science and Big Data, making the former leveraging the latter, as far as allowing a more efficient processing, compute, storage and access of the data, and as such, increasing the efficiency and viability of the various techniques within the Data Science field. With a practical perspective in mind, we did retrace some of the data transformation work we have conducted in large data centric organizations, as well as the specific use cases addressed within specific verticals, leveraging various Data Science techniques, on both the data management and machine learning and intelligent mining sides.

These use cases covered the areas of mobile applications, online advertising, security, healthcare and finance. This pragmatic approach did demonstrate the superiority of the results in some existing vertical industries, that have historically been fairly slow moving in terms of pushing new data analysis techniques, resulting in the creation of a platform for new revenue generating services and in some cases, to the emergence of a new generation of players, taking full advantage of the potential of recent Data Science models.

Xona Partners (Xona) is a boutique advisory services firm specialized in technology, media and telecommunications. Xona was founded in 2012 by a team of seasoned technologists and startup founders, managing directors in global ventures, and investment advisors. Drawing on its founders' cross functional expertise, Xona offers a unique multi-disciplinary integrative technology and investment advisory service to private equity and venture funds, technology corporations, as well as regulators and public sector organizations. We help our clients in pre-investment due diligence, post investment life-cycle management, and strategic technology management to develop new sources of revenue. The firm operates out of four regional hubs which include San Francisco, Paris, Dubai, and Singapore.

Xona Partners

www.xonapartners.com

advisors@xonapartners.com