ODBMS.org article – “On data analytics for finance. Interview with RavenPack’s Jason S.Cornez.”

> ***“Understanding human language remains a difficult problem. The challenges here are not only technical, but there is also a perception from popular culture that computers today perform at the level we see in science fiction. So there is a gap between what is expected and what is possible.”–Jason S.Cornez.***

I have interviewed **Jason S.Cornez**, *Chief Technology Officer, RavenPack.* Main topic of the interview is unstructured data analytics for finance.

RVZ

Q1. What is the business of RavenPack?

Jason S.Cornez: We specialize in the systematic analysis of unstructured data for finance. RavenPack Analytics transforms unstructured big data sets,such as traditional news and social media, into structured granular data and indicators to help financial services firms improve their performance. RavenPack addresses the challenges posed by the characteristics of Big Data – volume, variety, veracity and velocity – by converting unstructured content into a format that can be more effectively analyzed, manipulated and deployed in financial applications.

Q2. How is Deutsche Bank using RavenPack News Analytics as an overlay to a pairs trading strategy?

Jason S.Cornez: The profits and risks from trading stock

pairs are very much related to the type of information event which creates divergence. If divergence is caused by a piece of news related specifically to one constituent of the pair, there is a good chance that prices will diverge further. On the other hand, if divergence is caused by random price movements or a differential reaction to common information, convergence is more likely to follow after the initial divergence. To test the effects of news on a pairs trading strategy, Deutsche Bank used two aggregated indicators based on RavenPack's Big Data analytics derived from news and social media data measuring sentiment and media attention.
Specifically, using the two indicators, Deutsche Bank created a filter that would ignore trades where divergence was supported by negative sentiment and abnormal news volume.
Overall, Deutsche Bank finds that applying a news analytics overlay can help differentiate between "good" price divergence (which is likely to converge) and "bad" divergence. More importantly, such ability provides significant improvements to the performance of a traditional pairs trading strategy, especially by reducing divergence risk.

Q3. Who needs sentiment analytics in finance and why?

Jason S.Cornez: Sentiment analytics can help improve performance of trading strategies,reduce risk, and monitor compliance. Quantitative investors often subscribe to RavenPack Analytics granular data. This provides them with the ability to detect relevant, novel and unexpected events – be they corporate, macroeconomic or geopolitical -so they can enter new positions, or protect existing ones. These events, and the sentiment associated with them, help drive alpha generation as a novel factor in automated trading models.

Traditional Asset Managers, such as those managing hedge funds, mutual funds, pension funds and family offices may subscribe to RavenPack Indicators to help run portfolio optimization. The Indicators provide snapshots of sentiment and information density for an entity or instrument that can be used alongside fundamental or technical indicators to build portfolios with better risk/return profiles.

Brokerage and Market Makers can leverage RavenPack sentiment data to manage risk and generate trade ideas. They rely on RavenPack’s detection of relevant, novel and unexpected events – be they corporate, macroeconomic or geopolitical – to create circuit breakers protecting them from event risk.

Risk and Compliance Managers use RavenPack data to monitor accumulation of adverse sentiment or detect headline risk. The data help risk managers locate accumulations of risk and volatility, or changes in liquidity – either by aggregating sentiment, identifying event-driven regime shifts, or by creating alerts for when sentiment indicators reach extremes. As well, RavenPack event data also aids surveillance analysts to receive fewer false positives from market abuse alerts.

Finally, Professional and Academic researchers use RavenPack data to better understand how news and social media affect markets. They want to inform their clients how to find new sources of value and, hence, research and write about how quantitative investment managers find value in the data. RavenPack’s granular data is a great source of unique data for academics to enhance their published research – be it presenting a new way to use the data or controlling for news and social media in their work.

Q4. What are the main challenges and opportunities for Big Data analytics for financial markets?

Jason S.Cornez: Much of the work so far in Big Data analytics has been confined to structured data. These are sets of labeled and elementized values, such as what you might find in a traditional database table. Tools like Hadoop and Spark have helped to make structured big data analyitcs approachable.

RavenPack has always focused on unstructured data, primarily English-language text. Doing analytics here isn’t just about data mining, it requires more sophisticated processing for each document. Understanding human language remains a difficult problem. The challenges here are not only technical, but there is also a perception from popular culture that computers today perform at the level we see in science fiction. So there is a gap between what is expected and what is possible. One of our goals here is certainly to help make computers a little smarter.

Things start to get really interesting when you produce analytics by marrying structured data with unstructured data. A simple example could be a news story where an analyst expects mortgage rates to hit 4% by summer. It is certainly great if a computer understands that this is a story about interest rate guidance, but so much better if the computer is able to combine this with historical mortgage rates to know that the rates are currently rising, but still far below historical norms. As an industry, I don’t think too much has been done here yet, but that we’ll be seeing more activity here in the coming years.

Financial markets rely on information in order to be efficient. Big Data analytics promises to provide more information, and more types of information, faster than was previously possible. A more efficient market could help to level the playing field, as it were. And even if markets never become truly efficient, the financial industry sees that Big Data analytics can certainly help them. Several of these opportunities were addressed in the answer to the previous question.

Q5. What are your practical experience in building an infrastructure for Big Data Analytics of mostly unstructured text content, in realtime?

Jason S.Cornez: RavenPack has been processing Big Data since before Cloud Computing was a practical reality. We noticed that most competitors in the news analytics space were offering software solutions, whereas RavenPack has always been a service provider. We sell data, not software. As such we invested in our own infrastructure maintained at trusted hosting facilities. This was perhaps not the easiest or cheapest route, but it leads to compelling products that are relatively easy for a customer to adopt.

From the beginning, we’ve built a distributed system where collection, storage, classification, analytics, publication, and monitoring all run on distinct machines connected by a high-speed network. We learned virtualization technologies so that we could leverage our hardware investments more efficiently. We’ve been rigorous about maintaining a separation of concerns and establishing well-defined interfaces between our components. This not only makes our system robust, but it also allows us to choose the best technologies for each task.

In recent years, we’ve migrated to Cloud Computing and our early investments in distributed systems are really paying off. Most of our components work directly in the cloud and also scale without additional engineering work.

Q6. How do you manage to have a very low latency?

Jason S.Cornez: Low latency has always been a requirement of the system. Starting with low-latency, realtime processing in mind led to many of the architectural decisions that I mentioned above – especially about being distributed and being able to leverage big hardware. It’s painful to think about re-engineering an existing system that wasn’t designed with low latency in mind.

A specific observation is that storage, especially magnetic based storage, is far slower than CPU and also far slower than networking. So we have a heavily multi-threaded system where all storage tasks are delegated to background threads and the flow of data in the realtime system never needs to wait on a database.

Speaking of multi-threading, RavenPack performs various types of classification on each document. Many of these are independent and can be performed in parallel. As well, within a single document and single type of classification, many aspects work only on local information, such as a paragraph. This work can also be done in parallel. As more powerful, multi-core machines continue to appear, our system can continue to improve.

Of course, low latency really begins with good algorithms and good tools. We measure the system as a whole on a daily basis and we profile our code for both speed and space on a regular basis. At times, there is a trade-off between a feature and doing it feasibly. We often sacrifice a new feature until we can solve how to implement it without negatively impacting the performance of our system.

Q7. What are the main technological challenges you are currently facing?

Jason S.Cornez: There are many challenges ahead. Some of the obvious ones are about branching out from English into other languages, or from plain text to other media formats.

On the purely technical side, we see that cloud computing and big data are still very young fields. Cloud resources are much more ephemeral than those in a controlled, hosted environment. We must adapt software to work well in the face of disappearing machines and inaccessible resources. One example is startup time of a system. Traditionally, startup is a rare event and our servers run for a long time. But now that changes, and system startup is much more frequent and hence must be made more efficient. We are evolving rapidly in these areas right now.

Perhaps the biggest challenge remains the perception gap that I mentioned earlier. I’m very proud of the system we’ve built, but it remains possible for a human to find an entity or an event in a document that our system misses. I don’t think this problem will ever go away, but I’m confident RavenPack is making great strides here.

Q8. Why and how do you use Allegro Common Lisp?

Jason S.Cornez: RavenPack has been using Franz Allegro Common Lisp since we began. It is the primary language we use for analysis and classification of unstructured text. Common Lisp is an excellent language for both exploratory programming and high performance computing.

Common Lisp is a multi-paradigm language, or even a paradigm-neutral language. So the engineer has the flexibility to map from concept to code in the most natural way possible. Some concepts map naturally to an object-oriented design, others to a functional design, and other to an imperative design. The language naturally supports all of these so you never need to map from your concept into the philosophy of the language. And further, lisp is a programmable programming language, so as new paradigms come along, they can be added to the language by any developer. This is so easy and natural in Common Lisp that you often do it even when there is only a single use case in mind.

Common Lisp also shines for deploying and maintaining production software. Of course, it supports native OS threads, native machine compilation, and high performance garbage collection. But as well, you can attach to, inspect, modify and patch live systems.

Q9. What are the main lessons you learned so far?

Jason S.Cornez: It’s been a long and interesting journey, and nearly everything we know now has been learned along the way. One way I like to think about the main lessons learned is to consider what I believe to be the barriers that might make it difficult for a competitor or potential client to replicate what we’ve done.

A significant selling-point of our product that provides lots of value to our clients is our extensive historical archive of analytics. This of course is derived from our archive of content. The curation of such an archive is much harder than most people imagine. There is the minor issue of implementing the spec that the provider supplies. But the fun begins as you realize that the archive is incomplete and in multiple incompatible formats, some of them not documented at all. There are multiple timestamps, many with no timezone. The realtime feed looks different from the historical archive. The list goes on.

None of this is meant as a complaint about our content partners – this is the nature of things. And even having learned this lesson, there isn’t much we could have done differently. Of course, we now have a checklist of questions we give to any new content provider – and they often improve their offering as a result of working with us. But if we hear that incorporating someone’s content will be easy, we now know to take this with a grain of salt.

Qx Anything else you wish to add?

Jason S.Cornez: Thanks for this opportunity. I hope it has been helpful.

Jason S.Cornez, *Chief Technology Officer, RavenPack.*
Jason joined RavenPack in 2003 and is responsible for the design and implementation of the RavenPack software platform. He is a hands-on technology leader, with a consistent record of delivering break-through products. A Silicon Valley start-up veteran with 20 years of professional experience, Jason combines technical know-how with an understanding of business needs to turn vision into reality. Jason holds a Master’s Degree in Computer Science, along with undergraduate degrees

in Mathematics and EECS, from the Massachusetts Institute of Technology.