Heard on the Street – 4/18/2022

April 18, 2022 by Daniel Gutierrez Leave a Comment

Welcome to insideBIGDATA's "Heard on the Street" round-up column! In this regular feature, we highlight thought-leadership commentaries from members of the big data ecosystem. Each edition covers the trends of the day with compelling perspectives that can provide important insights to give you a competitive advantage in the marketplace. We invite submissions with a focus on our favored technology topics areas: big data, data science, machine learning, AI and deep learning. Enjoy!

Open Source Sabotage. Commentary by Mark Waggoner, Principal Engineer at LogRhythm

> *"The recent revelation of an open-source software package maintainer deliberately sabotaging the node-ipc package, which is part of the npm java package manager for the JavaScript programming language, has once again highlighted the dangers inherent in the Free Open Source Software (FOSS) supply chain. While we have seen several recent examples of malicious actors inserting their own code into software supply chains, like NotPetya in 2017, SolarWinds in 2020 and Microsoft Exchange Server in 2021, this appears to be if not the first instance of a maintainer deliberately sabotaging their own project, at least the most egregious in recent memory. This action adds a new potential threat actor to our risk assessments for using FOSS or FOSS-derived software. This scenario is very eye opening and frankly frightening. Even organizations who are following best practice guidelines put out by NIST and CISA would have no way of identifying this type of sabotage before it was active in their environment. Since this package was altered by the maintainer and then uploaded to the package manager through entirely valid workflows, it will have all the correct hashes and signatures that security professionals would check. In addition to that, it seems the maintainer also went out of his way to obfuscate their additions to the code base by using base64 encoding to make it even more difficult to identify by either automated means or human reading of the code. Today, with almost all software having at least some reliance on FOSS products, this action seriously increases the risks for any software company, developer, or even end user of these products. Unfortunately, even such things as "software manifests" and other mitigations as outlined in NIST 800-171 or the recent DHS report assessing supply chain risks would not have any impact on stopping this particular type of sabotage. However, having these risk awareness and mitigation policies in place should lead to quicker remediation once discovered.*

Why Data Backup and Recovery Need to Be Part of Your Zero Trust Security Program. Commentary by Ahsan Siddiqui, Director, Product Management at Arcserve

> *"The U.S. government recently announced that it is moving toward a Zero Trust approach to cybersecurity to dramatically reduce the risk of cyberattacks against the nation's digital infrastructure. The bottom line is that today's security is not secure.*

> *Organizations must assume bad actors will inevitably get in, and they must do everything to minimize their attack surface and protect their business-critical data from being damaged or destroyed. As part of this Zero Trust strategy, organizations must also be exceptionally vigilant around their data backup and recovery strategies. The concept of constantly verifying, continuously authenticating, and always logging who is going where and doing what should apply to regular operations and application usage. It should also apply to the data backup and recovery processes. For instance, it's critical to know who is initiating that backup and where they are backing up the data. The good news is that adopting Zero Trust for backup and recovery can mean extending the security controls that already exist within your environment. For example, applying multifactor authentication to your backup and recovery processes can go a long way toward establishing identity insurance and adding a greater level of protection to your organization. Immutable storage should also be part of any Zero Trust initiative. Immutability is when data is converted to a write-once, read many times format. Immutable storage safeguards data from malicious intent by continuously taking snapshots of that data every 90 seconds. Because the object store is immutable, you can quickly restore data even if someone tampers with it.*

How to protect your data science talent. Commentary by Dave Armlin, Field CTO, ChaosSearch

> *In today's competitive job market, organizations cannot afford to lose strong tech talent – especially in the IT, DevOps and data science departments where teams are already stretched thin for time and resources. Unfortunately, in the search for new tech talent, many business leaders are not actively looking for signs that their current employees are unhappy and at risk of leaving. A big part of this can stem from organizational structure and how these employees are seen by their colleagues. For example, it is a major red flag for data scientists to be viewed as a helpdesk-style IT resource for running queries and modeling data on behalf of other teams. Data science and tech talent can also be expected to spend a large portion of their time on mundane data cleaning or processing tasks versus putting their valuable skills to use by building models that will have an impact on the business. This often ties back to a larger problem with an organization's data architecture and outdated platforms that require an abundance of time spent on data prep and less time available for achieving powerful insights. With recent Data Science Day on April 6th, every organization should take the time to evaluate if their data scientists are wasting time on tasks they dread doing – like constantly helping other departments with requests or prepping data to be analyzed. Consider tools and new processes that will automate those mundane tasks and free data scientists to focus on what they're intended to do: provide better, more impactful insights that fuel business decisions. Otherwise, you could be at risk for losing coveted talent during a critical and competitive hiring war.*

Solving Data Engineer Burnout to Address the Talent Shortage. Commentary by Ori Rafael, CEO and co-founder, Upsolver

> *LinkedIn Recruiter recently showed over 200,000 open data engineering positions in the US, compared to roughly 30,000 employed data engineers. To compound the issue, modern data – big, complex, streaming – has both made data engineering increasingly important and essential, while also advancing the skills required in the role. While this is an industry-wide problem, it also makes the job a pressure cooker evidenced by an astounding 97% of data engineers reporting burnout at their workplace. Although the talent shortage is a macro problem, impacting the entire industry, companies can begin to address this by solving burnout at the micro level. This involves better training and tooling, specifically automating*

> *best practices into repetitive tasks. Not only will data engineers become more productive, but they will also be less likely to leave the field and the overall skill requirement for data engineers will be reduced, opening the market to a broader set of candidates.*

Professor Saiph Savage's ethical AI research makes an impact in rural communities. Commentary by Saiph Savage, Assistant Professor & Director Civic A.I. Lab, Khoury College of Computer Sciences, Northeastern University

> *The AI industry has created futuristic realities and continues to make technological advances in a number of industries and fields, such as healthcare, self-driving cars, and the stock market. However, there are systematic changes that still need to be made with regards to ethics and bias. For example, transparency of the labor conditions/work spaces of digital workers and unconscious biases among AI algorithms. Creating tools that empower digital workers who provide labels for the AI industry is absolutely essential. Quantifying workers' skill development, hourly wages, the integration of creativity, the amount of invisible labor and the amount of unfair evaluation received from employers, can be used with AI based tools to show progress and identify systematic biases. These AI based tools are meant to have an impact on sustainable development by providing ways through which we can directly quantify the labor conditions of digital workers, then do something about the problems that are quantified.*

Leveraging Synthetic Data To Develop Improved Customer Experiences. Commentary by Senthil Kumar Padmanabhan, Vice President, eBay

> *Synthetic data plays a critical role in product development by boosting developer productivity. One of the big impediments developers face while building products is the lack of quality real-world data to test their features confidently. A common and well-established idea proposed to address data issues is to create quality data in large quantities before executing the test cases and tear them down once done. Most organizations have well-defined APIs to create data; why not leverage them? In reality, this is easier said than done. It is easy to create monotonous or prosaic data in large quantities. However, it is nearly impossible for many organizations to create the millions of permutations and combinations of data required to execute the thousands of test cases required for a product launch. That is why we came up with the idea of taking a subset of production data and moving it to a test environment in a privacy-preserving way. The propensity of production data intervened with guaranteed privacy yields a high-quality testing environment. We saw that first hand at eBay. Previously for testing or feature development, teams relied on this laborious process of creating test data. It was tedious, time-consuming, sometimes manual, and even error-prone. When a part of the team was creating data, the rest were waiting on it, unable to make progress. They were in a blocked state, and it was a waste of time. A leadership principle that I follow and always tell my teams is, "Being blocked is a worse outcome than being wrong." When you are wrong, you at least did something, and it did not work out. But being blocked means you are waiting and not making progress. We want to avoid that at any cost. Now engineers always have access to high-quality data, which means they go less and less into the block mode, which is an enormous value for me. Secondly, our pass rate jumped to 95%, compared to only 70% in 2020. Flaky tests are a big frustration point in software development. Basically, engineers are chasing behind a false positive, which again is wasted time, as they are now blocked from doing something more important. Now that block is reduced. Furthermore, this is even more important when you are doing something at a platform layer like developer productivity since it impacts the whole company. Even a few minutes of saved time, multiply that with the number of engineers and the number of services. It has this compounding effect.*

Some sustainability solutions are computational, not behavioral. Commentary by Yuval Boger, Chief Marketing Officer, Classiq

> *Some sustainability problems can be addressed by encouraging different behavior: use less gas, recycle better, and prefer energy-efficient appliances. But major sustainability contributions can be made by solving thorny computational*

> *problems. For example, better simulation of molecular interaction could replace the Haber-Bosch process that produces ammonia yet requires nearly 2% of the global energy production. Better EV batteries would lower the cost while increasing the range of electric cars, making them even more attractive to consumers. Smarter sequencing and optimization of delivery routes would require less fuel while improving timeliness. What's holding us back? Some of these simulations are so complex that current classical computers cannot perform them and never will. Quantum computers offer hope as they can solve certain problems dramatically faster than classical computers: hours instead of millions of years when the problem is properly expressed and a sufficiently-advanced quantum computer is available. And, as an extra benefit, the quantum calculation process itself is much more energy-efficient than the classical process, providing hope that even the data centers of the future will be more energy-efficient.*

Model Failure Hindering ML Innovation. Commentary by Alessya Visnjic, CEO, WhyLabs

> *There is a huge blind spot holding back ML innovation in the enterprise. ML practitioners have nailed down model deployment, but they lack visibility into what happens once a model goes into production. This is resulting in an overwhelming rise in model failures, which ultimately is costing companies millions. The problem we face today is that the AI development process is far too similar to software development approaches. We treat software as code, and do not account for how data —which is the foundation of a ML model—affects the behavior of ML-powered software. Data is complex, highly dimensional, and its behavior is unpredictable. Therefore, the metrics we use to measure the performance of software simply can't adequately measure the performance of ML models. New metrics are needed to account for both infrastructure health and model performance, and also for the health of data that runs through the pipeline. The result: models are built with poor data, leading to model failure. Typical "Garbage in, garbage out" conundrum. If the data isn't healthy, the results are going to be subpar. AI models drift, become inaccurate, when the data coming in behaves differently from the training set or is not robust. In other words, models fail when they're not "ready" for the data they receive—and developers need the right tools to detect, observe, understand and explain why this is happening.*

Data as a revenue source? Integration for enterprise data collection & sales. Commentary by John Thielens, CTO of Cleo

> *Data is already incredibly valuable to enterprises seeking to optimize operations internally, and data very well could be just as valuable to external organizations. The exchange of data is already monetized by media companies and websites like Facebook. Why can't businesses in other industries do the same? Companies, especially those outside of the tech sphere, have long struggled to capture, standardize, and analyze the vast amounts of data needed to provide value to external organizations. That's because, to truly provide value through data, enterprises need to collect data on end-to-end operations – not just what occurs within their facilities. If organizations want to reap the full ROI of data collection and analytics, they need cloud-based integration technology that collects granular, holistic data from their ecosystem of partners and customers. This doesn't necessarily need to be consumer data. If supply chain and industrial enterprises can collect data on B2B transactions and the flow of goods, that could prove incredibly useful to research groups, financial institutions and more. Improved integrations that connect on-premise legacy systems to cloud-based applications can provide manufacturers, retailers, logistics companies and wholesalers the marketable data they need to maintain revenue in times of uncertainty. By connecting legacy systems to the cloud, supply chain organizations can quickly leverage cloud technology to gather external information and existing internal data. This could be the business-saving capability that keeps some small-to-mid-size enterprises afloat in the turbulent market of today.*

Unstructured Data: The Key to the Human Side of the Enterprise. Commentary by Kon Leong, CEO of ZL Technologies

> *Unstructured data shares one common characteristic: it's created by humans, for humans. This makes it incredibly powerful for understanding every element of the human side of the enterprise, including culture, performance, influence, expertise, and engagement. There's a reason Microsoft CEO Satya Nadella called a company's knowledge repository of communications "the most strategic database in the company." It's because employees share absolutely massive amounts of digital information and knowledge every single day, yet to this point it's been largely untapped. Fortunately, companies are now starting to leverage this intelligence goldmine. Using unstructured data such as emails, messages and files, organizations are now empowered to answer meaningful questions about their people, culture, and organization. For example: (i) Who are the most influential people in the organization?; (ii) Which employees are fulfilled with their work, and whom might we be at risk of losing?; (iii) How strong is company culture, and what obstacles are we facing when it comes to promoting diversity and inclusion? These questions about the "human side" of the enterprise are finally being brought into full view now with the help of people analytics.*

Using Low Code to Retain Top IT Talent. Commentary by Ed Macosky, Chief Innovation Officer at Boomi

> *A recent Gartner report found that only 29% of IT workers intend to stay at their current jobs. For employers looking to retain top IT talent, low code can be the solution. Using low code technologies, IT professionals can save time by not needing to write code from scratch. At the same time, low code can replace tedious, repetitive work and even eliminate a vast majority of debugging, which causes frustration. Low code will be responsible for 70% of application development by 2025, according to one leading analyst firm. This is because the possibilities for low code are endless – it can be used for designing user interfaces, API creation, application/data integration, process management, and more. Suddenly, using low code, IT professionals can spend their time on the most impactful or innovative work for themselves and the company, instead of on tedious tasks like cleaning up data. Companies need to focus on retaining top IT talent now more than ever, and low code can alleviate some of the stress on IT workers. By equipping IT professionals with time-saving tools, employers stand a fighting chance at attracting and retaining the IT professionals crucial to running their business.*

How Insurers Can Use Data to Foster Good Neighbors & Reduce Risk. Commentary by Benjamin Tuttle, Chief Technology Officer at Arturo.ai

> *Today, most homeowners get their insurance through a fairly manual process involving questionnaires and home visits and sometimes frustratingly slow processes. More to the point, homeowners themselves often don't fully understand that process, how to answer insurance-related questions, how they can reduce their home's risk or attain lower premiums. With the explosion of sensor and imagery technologies, there is a vast opportunity to modernize the fairly traditional insurance industry, especially at the neighborhood level. Instead of relying on outdated public records data or creating friction by asking the homeowner, "insurtechs" can apply AI to sensor technology to extract meaningful data. These insights can be as granular as the precise amount of rusting on a roof, how much debris is in the yard or how much of the trees overlap on the roof. These qualities paint a more accurate and current view of risk at a fraction of the cost, which can help reduce friction and improve policyholder experience. And at a neighborhood and portfolio level, this data has great potential to be transformative, enabling insurance companies to take on an advisory role with homeowners, providing home health checks and sending proactive and targeted communications to policyholders on how best to keep a home as risk-free as possible with simple actions like trimming trees or fencing a pool.*

What the Industry Needs to Build More Robust AI Models in the Future. Commentary by Yashar Behzadi, CEO and Founder of Synthesis AI

> *Self-driving vehicles made recent headlines as Ford launched a new autonomous driving division and Waymo sent driverless cars to San Francisco,*

> *prompting questions of if self-driving technology is ready for wide-scale deployment. As AI becomes increasingly incorporated into day-to-day life and grows in complexity and capabilities, the industry will face challenges maintaining and scaling these systems. AI systems built on static and manually labeled data will not hold up, as updating and managing these systems will be difficult and costly. Capturing edge cases and rare-events are difficult and required to ensure safe, predictable vehicle behavior. Hand labeled, manually collected data is labor-intensive, costly, prone to human error and bias, and presents consumer privacy concerns. Synthetic data, or computer-generated data that models reality, helps to eliminate these concerns. By training AI systems in simulated worlds, AV companies can quickly and cost-effectively build more robust models. Never before encountered rare events and edge cases can be tested to ensure proper vehicle responses. At a time of explosive AI growth, synthetic data can quickly deliver the data needed for more robust systems.*

The Environmental Impact of Data and How the IT Sector can Help. Commentary by Bennett Klein, Global Product Marketing Leader at Quest Software

> *According to a recent Gartner report, The IT sector totals 1.4% of global greenhouse gas emissions. What's more, data storage requirements are almost doubling annually — increasing the carbon footprint left behind by digital businesses. The environmental impact of data is real and growing, but there are ways for businesses to reduce their environmental impact. By using data to understand the nature and impact of your data, the efficiency of your data-driven operations, and ensuring you have an efficient and effective approach to data protection by using deduplication technology, organizations can surgically optimize their backup storage footprint, and significantly reduce the environmental impact of an organization's data requirements. Without a clear understanding of what data you have, how it is being managed, and how it impacts business, organizations will struggle to prioritize data processing and storage capabilities from an environmental perspective. Business efficiency and cost optimization can have a positive impact when it comes to your business' environmental footprint as well. Looking for ways that your business can optimize data governance, operations, data storage, retention and consumption can deliver a dual benefit of business integrity and environmental stewardship. With the right set of insight and capabilities in place, organizations do not have to choose between what is good for the business and what is good for the environment as natural synergies exist between the two.*

You have more data quality issues than you think. Commentary by Barr Moses, CEO and co-founder of Monte Carlo

> *The majority of data leaders I speak with intuitively understand their teams are challenged with maintaining high data quality and staving off broken dashboards. However, they drastically under estimate the extent of the issue even when they have strong internal data quality metrics. Our data observability platform has end-to-end access across hundreds of modern data stacks and millions of tables–our data shows the average organization will experience **one data issue per year for every 15 tables it hosts in its data warehouse**. There are many reasons data leaders lack visibility into the problem–from data drift to data engineer turnover to closed email based incident triage processes–but the solution is to scale monitoring either through writing more quality tests across your pipelines or automation.*

The Opportunity for Data Monetization. Commentary by Francis Wenzel, CEO and Co-Founder at TickSmith

> *From retail commerce stores to hedge funds and farmers, organizations big and small have the potential to become big data producers. For the approximately 65% of firms missing out on this untapped opportunity, the problem is they often let valuable data go to waste because they don't know where to begin. Currently, firms that are taking advantage of this fast approaching 'data economy' are corporations spanning a wide range of industries acting on the realization that in-house data sets represent a valuable external asset to their bottom line. As the world becomes more 'data-driven', the opportunity grows in parallel for companies to leverage unused data sets to generate new revenue streams, improve efficiency and increase sales. If more*

> *organizations embrace the notion that all data is valuable in some shape or form, and can organize, harness and commercialize this asset, they will capitalize on new revenue streams and grow against the competition.*

Sign up for the free insideBIGDATA newsletter.

Join us on Twitter: @InsideBigData1 – https://twitter.com/InsideBigData1

Related Posts

The insideBIGDATA IMPACT 50 List for Q4 2021

The insideBIGDATA IMPACT 50 List for Q3 2021

The insideBIGDATA IMPACT 50 List for Q4 2020

The insideBIGDATA IMPACT 50 List for Q2 2021

The insideBIGDATA IMPACT 50 List for Q3 2020

Filed Under: AI Deep Learning, Analytics, Big Data, Big Data Hardware, Big Data Services, Big Data Software, Cloud, Data Science, Data Storage, Database, Featured, Google News Feed, Machine Learning, News / Analysis Tagged With: AI, Big Data, data science, Deep Learning, Machine Learning, Weekly Newsletter Articles

Leave a Comment

Name *

Email *

Website

Notify me of follow-up comments by email.

Notify me of new posts by email.

Post Comment

Resource Links: