Design the Data Lake Configuration to address the Technology Requirement for Stitch Fix

Group 2: Claire Wu (jw3644), Mandy Chan (mc4528),
Sarah Garman (sag2213), Teressa Newton (ten2111)

Columbia University School of Professional Studies Applied Analytics
APAN 5400: Modern Database Architecture
Requirements Document Assignment

Authors' Note
This paper is being submitted on July 31, 2018, for the completion of APANPS5400 Requirements Document - Final Version of Technology Section

Executive Summary

The following is a technology requirement proposal to improve Stitch Fix's recommendation algorithm accuracy through the introduction of a new celebrity quiz and incorporating social media APIs. The technical team reviewed various system that is capable of scaling large and processing real-time data in formats such image and video. As a result, we propose MongoDB and Neoj4 as Stitch Fix's database engine. Additionally, after careful consideration of the various tools, we developed a primary and a secondary data lake architecture recommendation. The specifics of the data lake components are discussed in great detail along with recovery/continuity of business should an emergency occur, and data governance.

Business Requirements Review

Our database recommendations will need to conform with the data volume and storage requirements as outlined in the business requirements: new quiz data is predicted to be 15% of Stitch Fix's (SF) current data volume and the volume of celebrity-related data from social media sources is estimated to be 10% of their total data volume.

There are two main sources of data we need to incorporate into Stitch Fix's existing data architecture: the celebrity quiz and the social media data. This new information will need to be integrated into Stitch Fix's existing recommendation system and algorithm development program. Currently, they use several tools that integrate within the Hadoop ecosystem: S3, Hive, and Docker (Krawczyk, 2017). They have also built their own APIs internally, so our integration with the celebrity information may be developed in-house.

After consulting with the business team, the quiz questions contain both structured and unstructured data (users will be prompted to enter in their own celebrity choices), and the social media API will contain unstructured and streaming data. Structured data can be handled well by a traditional relational database such as Oracle, but it can also be stored in a NoSQL database with the social media information. Our recommendations below will go into detail on the specific NoSQL databases that would support the business recommendation.

The Nature of Stitch Fix Data

There are three key concepts to evaluate the nature of SF data, the 3Vs of big data (Gewirtz, 2018).

1. **Volume:** the existing quantity of data in Stitch Fix is huge. With the increasing of its user base, product categories, and business segments, the collection of data will increase exponentially (Stitch Fix, 2017).
2. **Velocity:** one of the features of SF data is real-time updating. This can be explained by its business nature. SF is learning customers preferences and trending styles by gathering questionnaire answers and social media research (mainly through Pinterest). Therefore, its styling recommendations and related data are dynamically changing (Stitch Fix, 2017).
3. **Variety:** SF data can be separated into a structured data type and an unstructured data type. Structured data include customer fit feedback and purchase histories. Unstructured data include photographic and textual data, such as inventory style photos, Pinterest boards - 'pinned' images, and the vast amount of written feedback and request notes from clients (Stitch Fix, 2017).

Database Engines Evaluations

MongoDB

Benefits: Three major attractive features of MongoDB that are appropriate for this project are their flexible schema and social media analytics capability. First, MongoDB is a non-relational and schema-free database, giving us the greatest flexibility to store any types of data without a previous structure defined. SF needs this type of platform to handle their item database that is becoming increasingly large and diverse (Data Flair, 2018). Second, MongoDB is widely used for real-time analytics in various sectors. This is extremely useful in terms of analyzing real-time celebrity fashion trends to improve SF's recommendation system (Venkatraman, 2017). Third, MongoDB is a good platform for storing video and images. These features combined with MongoDB's high-speed capability makes it an attractive platform for SF (Kanoje et al., 2015).

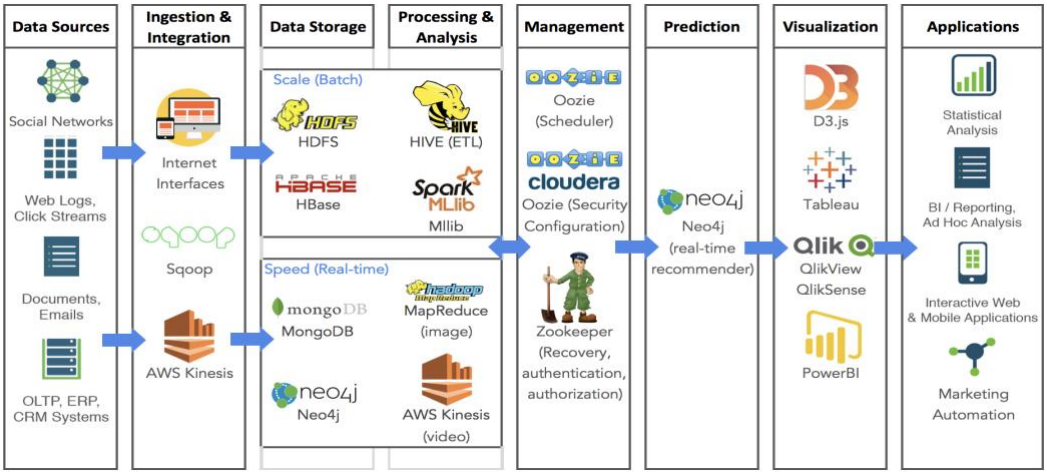
Risks: There are four limitations present in MongoDB. First, unlike a relational database, MongoDB does not support joins. Joins can be manually coded, but it is time consuming. Second, since MongoDB lacks the join capability, problems such as data redundancy arise, creating unnecessary usage of memory. Third, document size cannot exceed 16 MB. Finally, document nesting cannot exceed 100 levels (Data Flair, 2018).

Neo4j

Benefits: Neo4j, as a graph database tool, is a NoSQL database engine for relationship modeling, powering the real-time recommendations. Since SF is a matchmaker that connects its users with suitable styles, a high-quality recommendation system is critical to SF's business. Neo4j is an ideal tool to create user item relationship graphs based on affinity. It provides for deep queries to create recommendations for exploring multiple levels of affinity (Kumaran, 2016). Additionally, it provides a naturally adaptive schema optional data model, offering the flexibility to extend for new attributes (Boyd, 2016). For SF, adding the new variable, celebrity, will create new nodes/ properties/ relationships upon the existing database. Since no schema migrations are needed, we can just add the new features.

Risks: One of the risks is that SF cannot shard Neo4j, which means the whole dataset should be stored on only one server (Rund, 2017). Plus, it has a storage limitation for 34 billion nodes and relationships, and 274 billions of attributes for the current version (Luu, 2012).

Data Lake Components



Data Lake Architecture - Primary Recommendation

By leveraging the celebrity-related data from social media platforms and incorporating the insights to the newly launched celebrity quiz, the business objective of this project is to learn the consumers' preferences better and ultimately decrease the return rate of SF boxes. Hence, the primary recommendation is mainly designed for the consideration of processing and analyzing the unstructured data of celebrities and fashion influencers. Typically, those data include a large amount of images and videos from dynamically changing social media platforms. Therefore, the decision to utilize MapReduce for image analysis (MathWorks, 2018) and AWS Kinesis for video ingestion and analysis (AWS, 2017) allows the team to effectively turn the unstructured data into trends, fashion styles, and mainstream preferences in a real-time manner. Then the celebrity quiz could be iterated to learn the changing customer preferences. From data storage perspective, the integration of Neo4j provides SF the capabilities of NoSQL data storage, real-time recommendations, and adaptive schema optional modeling. And MongoDB can support SF with image and video storage, flexible schema, and social media analysis capabilities.

On top of the streaming data framework, the team leverage HDFS to support MapReduce with very high bandwidth (Eric, 2012) and HBase to support time-saving data reading and online analytical operations (Guru, n.d.). The integration of HIVE provides SF the ability to analyze data by writing SQL like queries - HiveQL, which largely simplifies working with large amounts of data (Ritika, 2017). Plus, utilizing Spark MLlib could drastically improve the machine learning efficiency of SF for structured data - 100 times faster than Hadoop for large-scale data processing (Joseph, 2016).

Data Lake Architecture - Secondary Recommendation

The secondary recommendation proves a slight modification of the current SF infrastructure to extract, transform, and load data meeting data needs with less emphasis on the real-time data. Hence, this option considers the scalability of the data as well as flexibility retrieving information. The distributed architecture allows one way to scale information with HDFS as the standard. However, there are additional tools for adequate integration within the data infrastructure. Spark provides a way to integrate streaming, machine learning, and graphing using various programming languages which increases accessibility but as the consequence of real-time analysis (Spark, 2018). An additional consideration is integrating a cloud-based application that can ease the end-user experience separating storage from processing responsibilities (Knight, 2017). Apache Cassandra provides scalability by dividing data from the processioning (Apache Cassandra, 2018). This recommendation requires the continued use of YARN for resource management. The visualization options remain unchanged. However, the final outputs from the application will minimal significant change.

Recovery/Continuity of Business

One hour of downtime can cost a midsize company such as Stitch Fix up to $75,000 in economic and operational losses (Shaw, 2018). Therefore, business continuity and disaster recovery strategies play a crucial role in the case of an unplanned event, such as cyber attacks, human error, or natural disaster. Risk analysis and strategic planning should be performed ahead of time and reviewed constantly.

In terms of risk mitigation, maintenance is an important element. Thus, it is crucial that all systems are constantly up-to-date and data backup are performed daily. In addition, for safety measure, Stitch Fix will need to establish a disaster recovery team and compile an updated employee and external stakeholder contact list.

Contrarily, to achieve operational recovery in response to an emergency, options to consider in the disaster recovery plan include data replication, initiating alternate network routes, working remotely, and falling over to a cloud-based service (Rouse, 2017).

Another element to consider that is specific to the various Hadoop tools we plan to adopt is to hire a Hadoop admin. While Hadoop is powerful, it comes with extreme complexity in terms of its ecosystem setup and maintenance. A Hadoop admin will ensure cluster maintenance, manage resource and security, and provide backup and recovery task (BigData, 2018). Additionally, Apache Zookeeper is included in our data lake configuration to enable synchronization across Apache Hadoop cluster. For this reason, Zookeeper is great tool for data recovery, authentication, and authorization (Rabczak, 2016).

Data Governance

Currently, Stitch Fix does not have a Chief Data Officer (Stitch Fix, n.d.). As such, we propose that the duties of data governance reside under the Chief Technology Officer's purview. The purpose of having a strong data governance plan is to ensure the quality of the new data we will be adding to the Stitch Fix's existing algorithms. Based on our data lake configuration, the following tools may be useful to supplement Stitch Fix's data governance strategy. Additionally, all three suggestions promote the collaboration between business and IT due to the intuitive nature of the audit and compliance dashboards.

Cloudera Navigator: One of the best tools for integrating with Hadoop. It may be worthwhile to invest in this tool as we are using tools from the Hadoop ecosystem in our data lake architecture. Navigator features auditing, data and metadata tracking, and policy management (Cloudera, n.d.).

SAS Data Governance: The strong point of SAS Data Governance is its business data glossary, which is a "powerful, interactive repository of terms and definitions that make it easier to flag and fix issues" (SAS, n.d.).

IBM InfoSphere Information Governance Catalog: One of the major benefits of this product is that it is web-based and comes with storage in the cloud.

References

1. Alex, M. & Noah, D. (2017, May). HDFS vs. HBase : All you need to know. KDnuggets. Retrieved from https://www.kdnuggets.com/2017/05/hdfs-hbase-need-know.html
2. AWS. (2017, November 29). Introducing Amazon Kinesis Video Streams. AWS. Retrieved from https://aws.amazon.com/about-aws/whats-new/2017/11/introducing-amazon-kinesis-video-streams/
3. BigData. (2018). Hadoop Administration and Maintenance. BigData. Retrieved from http://www.hadoopadmin.co.in/hadoop-administration-and-maintenance/
4. Bista, N. (2018, February 28). HDFS vs. HBase - Which One Is Better (Infographics). EDUCBA. Retrieved from https://www.educba.com/hdfs-vs-hbase/
5. Boyd, R. (2016, March 14). Intro to Graph Databases Episode #2 - Properties of Graph DBs & Use Cases. YouTube. Retrieved from https://www.youtube.com/watch?v=-dCeFEqDkUI&t=555s
6. Cloudera. (n.d.). Big data meets data governance. Cloudera. Retrieved from https://www.cloudera.com/products/product-components/cloudera-navigator.html
7. Craig, S. (2017, January 13). AWS Summit 2017 - Driving Business Outcomes with a Modern Data Architecture. YouTube. Retrieved from https://www.youtube.com/watch?v=6JeU-gsxHXk
8. Data Flair. (2017, May 1). Apache Spark Ecosystem – Complete Spark Components Guide. Data Flair. Retrieved from https://data-flair.training/blogs/apache-spark-ecosystem-components/
9. Data Flair. (2018, April 17). Advantages of MongoDB | Disadvantages of MongoDB. Data Flair. Retrieved from https://data-flair.training/blogs/advantages-of-mongodb/
10. Eric, B. (2012, July 25). Thinking about the HDFS vs. Other Storage Technologies. Hortonworks. Retrieved from https://hortonworks.com/blog/thinking-about-the-hdfs-vs-other-storage-technologies/
11. Gewirtz, D. (2018, March 21). Volume, Velocity, and Variety: Understanding the three V's of big data. ZDNet. Retrieved from https://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/
12. Guru99. (n.d.) HBase: Limitations, Advantage & Problems. Guru99. Retrieved from https://www.guru99.com/hbase-limitations-advantage-problems.html
13. IBM InfoSphere Information Governance Catalog. Retrieved from https://www.ibm.com/us-en/marketplace/information-governance-catalog
14. Joseph, B. et al. (2016, February 11). Why you should use Spark for machine learning. InfoWorld. Retrieved from https://www.infoworld.com/article/3031690/analytics/why-you-should-use-spark-for-machine-learning.html
15. Kanoje, S. et al. (2015). Using MongoDB for Social Networking Website: Deciphering the Pros and Cons. EEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems. Retrieved from https://arxiv.org/pdf/1503.06548.pdf
16. Krawczyk, S. (2017, January 23). Scaling Data Science: Slides from #DDTX17. Multithreaded. Retrieved from https://multithreaded.stitchfix.com/blog/2017/01/23/scaling-ds-at-sf-slides-from-ddtexas/.
17. Kumaran, P. (2016). Data Science Foundations: Choosing the Right Database. Lynda.com. Retrieved from https://www.lynda.com/MyPlaylist/Watch/17364682/754811?autoplay=true
18. Luu, J.W. (2012, June 12). Pros and cons about graph databases and especially Neo4j. Google Groups. Retrieved from https://groups.google.com/forum/#!topic/neo4j/mts6H9Py-2I
19. Mailajalan, A. & Data, N. (2017, May). HDFS vs. HBase: All you need to know. KDnuggets. Retrieved from https://www.kdnuggets.com/2017/05/hdfs-hbase-need-know.html
20. Masters, G. (2017, January 4). MongoDB Databases Under Attack Worldwide. Retrieved from https://www.scmagazine.com/mongodb-databases-under-attack-worldwide/article/629601/

21. MathWorks. (2018). Process Large Set of Images Using MapReduce Framework and Hadoop. MathWorks. Retrieved from https://www.mathworks.com/help/images/process-large-set-of-images-using-mapreduce-framework-and-hadoop.html
22. MongoDB. (n.d.). NoSQL Databases Explained. Retrieved from https://www.mongodb.com/nosql-explained
23. Rabczak, K. (2016, December 12). Data Synchronization – Part 1: Service Discovery and Leader Election. TheCookieZen Blog. Retrieved from http://thecookiezen.com/blog/2016/12/12/data-synchronization-in-distributed-environment-part-1/
24. Rakesh, R. & Michael, H. (2017, January 24). Hardening Apache ZooKeeper Security: SASL Quorum Peer Mutual Authentication and Authorization. Cloudera. Retrieved from https://blog.cloudera.com/blog/2017/01/hardening-apache-zookeeper-security-sasl-quorum-peer-mutual-authentication-and-authorization/
25. Ritika, P. (2017, October 23). What are the advantages of Apache Hive. Quora. Retrieved from https://www.quora.com/What-are-the-advantages-of-Apache-Hive
26. Rouse, M. (2017, July). Business Continuity and Disaster Recovery (BCDR). Tech Target. Retrieved from https://searchdisasterrecovery.techtarget.com/definition/Business-Continuity-and-Disaster-Recovery-BCDR
27. Rund, B. (2017, March 14). The Good, The Bad, and the Hype about Graph Databases for MDM. TDWI. Retrieved from https://tdwi.org/articles/2017/03/14/good-bad-and-hype-about-graph-databases-for-mdm.aspx
28. SAS. (n.d.). SAS Data Governance. SAS. Retrieved from https://www.sas.com/en_us/software/data-governance.html
29. Sharad, V. (2017, November 2). What is unstructured data and how to process it on Hadoop. Ovaledge. Retrieved from https://ovaledge.com/processing-unstructured-data/
30. Shaw, K. (2018, January 23). What is disaster recovery? How to ensure business continuity. NetworkWorld. Retrieved from https://www.networkworld.com/article/3248969/data-center/what-is-disaster-recovery-how-to-ensure-business-continuity.html
31. Shubham, S. (2017, September 6). How do I import unstructured data to Hadoop. Quora. Retrieved from https://www.quora.com/How-do-I-import-unstructured-data-to-Hadoop
32. Sridhar, V. & Christopher, C. (2015). Hadoop Image Processing Framework. IEEE. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7207264
33. Stitch Fix. (2017). Stitch Fix Algorithm Tour. Stitch Fix. Retrieved from https://algorithms-tour.stitchfix.com/#data-platform
34. Stitch Fix. (n.d.). Stitch Fix Resources. Stitch Fix. Retrieved from https://newsroom.stitchfix.com/resources/#Resources--team
35. Subramaniyaswamy, V. et al. (2015, May 8). Unstructured Data Analysis on Big Data using MapReduce. Procedia Computer Science. Retrieved from https://ac.els-cdn.com/S1877050915005165/1-s2.0-S1877050915005165-main.pdf?_tid=912f689d-3fb3-4a32-8f87-97b9a218d888&acdnat=1532445436_d8947e44bfda7163f3bac3dcb151530b
36. Venkatraman, P. (2017, March 21). MongoDB For Real Time Analytics. Analytics Training. Retrieved from https://analyticstraining.com/mongodb-real-time-analytics/

Image Sources

https://www.google.com/search?biw=1215&bih=953&tbm=isch&sa=1&ei=etZcW8PeBfKh_QbJjo3oDQ&q=powerBI+logo&oq=powerBI+logo&gs_l=img.3..0j0i10k1l7j0i5i10i30k1j0i10i24k1.115069.118630.0.119433.7.7.0.0.0.0.56.370.7.7.0....0...1c.1.64.img..0.7.368...0i67k1j0i7i30k1j0i7i10i30k1.0.N0wMycAPpT4#imgdii=92A0q7_AnP3h9M:&imgrc=eaxcLLtX4bt-bM:

https://www.google.com/search?biw=1215&bih=953&tbm=isch&sa=1&ei=tdVcW9WtMMnp_Qb8_aZ4&q=tableau+logo&oq=tableau+logo&gs_l=img.3..0l9.94143.97503.0.97758.12.11.0.1.1.0.121.609.10j1.11.0....0...1c.1.64.img..0.12.611...0i67k1.0.Xy_fPimsj_Q#imgrc=V_UrdaHV54072M:

https://www.google.com/search?biw=1215&bih=953&tbm=isch&sa=1&ei=sNVcW_KzHOPD_Qaq5Y0Y&q=D3.JS+LOGO&oq=D3.JS+LOGO&gs_l=img.3..0.3690.4380.0.4645.5.5.0.0.0.0.54.253.5.5.0....0...1c.1.64.img..0.5.252...0i30k1j0i24k1.0.A_5lGzs5TDQ#imgrc=YGjQfZE1oviC2M:

https://www.google.com/search?q=internet+interfaces+icon&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjQ-83DtMLcAhVmhuAKHfo4DrQQ_AUICigB&biw=840&bih=953#imgrc=2pB8SMSDZh4FRM:

https://www.google.com/search?biw=840&bih=953&tbm=isch&sa=1&ei=p8tcW9iqH86b5gKWkJHABg&q=zookeeper&oq=zookeeper&gs_l=img.3..0l10.191708.195302.0.195465.15.9.3.3.3.0.83.502.9.9.0....0...1c.1.64.img..0.15.518...0i67k1j0i10i24k1.0.CYIP-QGQ79E#imgrc=urQPRMVASpc7JM:

https://www.google.com/search?biw=840&bih=953&tbm=isch&sa=1&ei=ZbhcW5qxMs2c_Qb7x6OICg&q=HDFS+ICON&oq=HDFS+ICON&gs_l=img.3..0l2j0i5i30k1.1231324.1234275.0.1234512.9.9.0.0.0.0.94.546.9.9.0....0...1c.1.64.img..0.9.544...0i67k1j0i10k1j0i8i30k1j0i24k1.0.qVRH3WjqMpg#imgrc=HN2eIUMrSfeJIM:

https://www.google.com/search?biw=840&bih=953&tbm=isch&sa=1&ei=OL1cW4D0MOij_QaTx6SIDQ&q=hbase+logo&oq=HBASE&gs_l=img.1.1.0j0i67k1l3j0j0i67k1j0l4.244075.244497.0.246379.4.4.0.0.0.0.70.175.3.3.0....0...1c.1.64.img..1.3.174....0.HS8HpvBxvD8#imgrc=6WjrFg4_SaRE5M:

https://www.google.com/search?q=neo4j&source=lnms&tbm=isch&sa=X&ved=0ahUKEwj0kLa5wMLcAhXSnuAKHUS_A7wQ_AUIDCgD&biw=840&bih=953#imgrc=8gxcCFf1EpHqYM:

https://www.google.com/search?biw=840&bih=953&tbm=isch&sa=1&ei=H79cW7O5HIu35gKiwK2QBw&q=MLlib+logo&oq=MLlib+logo&gs_l=img.3..0j0i5i30k1.679327.684592.0.684842.14.11.3.0.0.0.167.675.9j1.10.0....0...1c.1.64.img..1.13.681.0..0i67k1j0i24k1j0i10i24k1.0.3jmam1lksGs#imgrc=t9tFUVU_vk0Y-M:

https://www.google.com/search?biw=840&bih=953&tbm=isch&sa=1&ei=zcFcW9uHBYTK_Qa055Eg&q=hive&oq=hive&gs_l=img.3..0i67k1j0l3j0i67k1j0l3j0i67k1l2.182629.182889.0.183087.4.4.0.0.0.0.91.272.4.4.0....0...1c.1.64.img..0.4.271....0.FNKu7blwm8Q#imgrc=S2eCkw9pyosNFM:

https://www.google.com/search?biw=840&bih=953&tbm=isch&sa=1&ei=YcNcW73sB_Ca_Qb8_5_wCQ&q=mapreduce+logo&oq=mapreduce+logo&gs_l=img.3..0j0i5i30k1.2268.2743.0.2931.5.5.0.0.0.0.61.273.5.5.0....0...1c.1.64.img..0.5.273...0i67k1j0i24k1.0.uWveoVZHrQU#imgrc=EEYO_bP_C_invM:

https://www.google.com/search?q=spark&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjZrMiYxsLcAhUlTd8KHWd5CHAQ_AUICigB&biw=840&bih=953#imgdii=67M9HNGuMvE08M:&imgrc=BWx6AhPbFmV70M:

https://www.google.com/search?biw=840&bih=953&tbm=isch&sa=1&ei=oMtcW6eSGOeH5wK8t66IBw&q=cloudera&oq=cloudera&gs_l=img.3..0l10.4187.6315.0.6487.8.8.0.0.0.0.122.555.7j1.8.0....0...1c.1.64.img..0.8.553...0i67k1j0i10k1.0.BTRXemEYsys#imgrc=Gz-54s9tS28n3M:

https://www.google.com/search?q=data+lake+architecture&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjAzfecmMLcAhUKh-AKHVUWD40Q_AUICigB&biw=1018&bih=953&dpr=2#imgrc=PkcpfKV0al7ZEM:

https://www.google.com/search?biw=840&bih=953&tbm=isch&sa=1&ei=VY9cW9GNNaSp_QbRl52ADw&q=third-party+icon&oq=third-party+icon&gs_l=img.3..0i7i30k1l7j0i30k1l2.7923127.7926652.0.7927669.11.11.0.0.0.0.123.615.9j1.10.0....0...1c.1.64.img..2.9.550...0j0i10i67k1j0i10k1.0.CdKdDw7vqPU#imgrc=-kqg3Gt8OvUKHM:

https://www.google.com/search?biw=929&bih=953&tbm=isch&sa=1&ei=ebJcW87wOKme_QawsLSwAg&q=Amazon+Kinesis&oq=Amazon+Kinesis&gs_l=img.3..0l8j0i30k1l2.39215.39215.0.39783.1.1.0.0.0.0.55.55.1.1.0....0...1c.2.64.img..0.1.55....0.rssWZDCEiuQ#imgdii=63D0qFrr0wqXOM:&imgrc=OtIBl4Fb81NITM:

APPENDIX

Sample quiz questions from consultation with the business team (verbatim) determined that the quiz questions contain both unstructured and structured data types:

"To make your styling journey with Stitch Fix even more exciting, we propose you to share with us something on celebrities and public figures who inspire your style. Jump in!" (* feel free to skip if this is not your thing)

1. Pick up a celebrity, who inspires your style most! (unstructured text)
2. Do you believe style is one of the attributes of his/her success?
    a. Yes
    b. No
3. What is so great about his/her style? (unstructured text)
4. What of the following fits your thinking (structured data):
    a. I love his/her style as he/she is a wonderful personality
    b. This is more about style, I'm not sure what kind of person he/she is
    c. I actually don't approve her/his deeds, but the style is a killer
5. I want to be like him/her (a scale from 1 to 10, structured data)"