Report 1: Student Internship Profile

Please complete all fields.

Name:	**Rahul Ethiraj**
Company:	A9.com
Department/Team:	Advertising Technology
Project Name:	Amazon Experimentation Framework & metrics- AXF

1. Describe the company at you are interning:

- The A9 Search Technology group provides the advanced search technologies that power Amazon product search.
- The mission is to provide a high-availability, low-latency search service that makes it easy for searchers to find the products they want.
- We are also making these same advanced search technologies available as a self-service search service that can be used by groups throughout Amazon for other purposes, such as searching online help or customer support information.

2. Please summarize your responsibilities:

- I currently work as a developer and tester for a software feature called Ad Runway Metrics pipeline with statistical metrics in Advertising Technology, Amazon group.
- I am responsible for integrating this new feature with the existing Amazon Ad Exchange (AAX) console. Integration includes working in technology platforms such as Big-data, Spark, PySpark, and other internal technologies for Amazon.
- One of my duties include taking care of team’s lab devices and work in designing topology and building connections based on requirements from the team members for testing a new feature. This includes determining hardware and software compatibility among devices using internal regression test tools.
- I need to design a new pipeline which should parse raw AAX logs and calculate metrics relating to Advertising technologies such as Bids, Impressions, Clicks and Punts.
- I am accountable to automate as I develop and let the machines do testing for us on a continuous basis.

3. Describe the team and/or role that you have been assigned:

- Currently, my team is relying on ADP cluster for its data processing but with increased load on cluster and increased number of use cases, ADP is unable to deal with the issues of scale. Also, there is very limited to no support for spark on ADP.
- We as a team plan to leverage the Spark framework for abstracting much of the complexity of dealing with coding in the Map/Reduce paradigm.
- Further, the team plans on using an amazon framework called Aegis to deal with the cluster operations (i.e., creating cluster, code deployment, running code, and terminating clusters).

- My team is called Newton Experimentation and plays a crucial role in calculating various Advertising metrics that will be used to evaluate various experiments conducted.
- The team consists of total 4 members including the manager. It has two senior technical leaders, acting as a go to person for any technical related issues and one team lead.
- The team runs on Agile software model and each member will be a scrum master in rotation. Each sprint is of two weeks length and a team member is expected to complete the assigned work in the user story keeping track of the hours/effort spent for each story. Team meetings are conducted once in each sprint phase.
- It uses various internal tools such as Amazon Chime for internal communication among employees, live video meetings, Cisco AnyConnect for VPN services.
- My team uses web-based tool to manage resources and test beds within the team. This tool is homegrown, simple to use and is integrated with AAX. The tool provides controlled access to AWS clusters. My task is to maintain the tool and keep an eye to support the team's lab repository.
- As an intern my primary job is to ramp up and learn as much as I can about the AWS clusters and big data so that I can start working on extracting the before mentioned metrics from raw AAX logs.
- My role in the team is both developer and tester crucially working on this feature and push it to the production code, after identifying any possible defects.
- I contribute in adding extra two statistical metrics such as p-value, and confidence interval and then have it verified by a senior team member before pushing it into the master production code.

4. Describe project(s) you will be undertaking:

- The goal is to populate metrics report for every hourly logs generated by AAX server within one hour, and make PROD level old Ad Runway Metrics pipeline that also calculates statistical metrics such as p-value and confidence intervals.
- Experiment metrics was calculated as part of experimentation and reports were available on AAX console. This pipeline was stopped in past due to cost reasons. This task aims in recreating the metrics flow pipeline.
- The goal is to populate metrics report for every hourly logs generated by AAX server within one hour. The old pipeline is no more applicable. However, the metrics format remains same.

There are a variety of reasons to explore using Spark on our EMR clusters. These include:

1. Currently we rely on ADP cluster for our jobs processing which is subject to cluster utilization and thus we have little to no control on their processing SLA's. We would be able to gain better control on the jobs processing SLA.
2. It also opens up possibility of doing ad-hoc analysis on top of EMR with no dependency on ADP team for adding and making libraries required for processing available on ADP cluster. We own, deploy and test our own libraries and processing end to end.
3. Moving to EMR is a positive direction forward in bringing in more tools and being able to accommodate the needs of the larger team.