**Annai Systems BioComputeFarmTM Proof of Concept Agreement**

**Customer Information**

Customer Name Station X

Address 185 Berry Street, Suite 5503

City San Francisco

State / Zip code CA 94107

POC project manager

POC technical manager Sandeep Sanga, PhD

POC IT contact Adin Stein (adin@stationxinc.com)

POC business manager

**Annai Systems**

Account Executive: Scott McIntee

Phone: (760) 213-1278

Email: scottm@annaisystems.com

Responsible Engineer: Mike Ainsworth

Phone: (415) 385-5705

Email: mainsworth@annaisystems.com

**POC description:**

Station X is currently evaluating cloud and managed compute solutions. Station X will use this compute resource to execute their variant calling pipeline to “normalize” TCGA DNA Sequencing files. Station X will also use Annai Systems’ BCF (BioCompute Farm) in conjunction with GTFuse to perform analyses that require the ability to select regions of a given BAM file for subsequent analysis.

Station X to be given access to Annai’s BCF to create a virtual machine to facilitate the analysis described above, for 14 days.

**The customer and environment**

Station X has a data analysis platform called GenePool that is used for cohort scale biomarker discovery. The platform is designed to work with RNA-Seq and DNA sequencing data. Station X is working with the TCGA data set to overcome some of the inherent sources of variability in the data, thus making it easier for researchers and clinicians to use this data set as part of their workflow in GenePool. Presumably, Station X will need to house this data set (and others like it that they may make available to their customers) in a cost effective, easy-to-use environment that is integrated with GenePool in a streamlined manner.

* + What tools is the cxustomer going to use
    - (1. System 2. Analysis) tools will be pulled from the library and which one will be loaded by the customer.
  + What is the ideal config
    - Software features sought (additional)
  + What OS will the client be running on
  + Customer IT requirements:
    - Will your IT team need to open a port?
    - Will be using SSH, VPN Tunnel, HTTPS://
    - What is bandwidth available to connect customer systems to BCF
  + Security policy requirements
    - Customer will receive a static IP to Annai’s bastion host. How will this affect internal security policies?

**The problems to be solved**

1. Provide compute capability and related infrastructure in a manner that eliminates the need to download BAM files into a local compute environment.
2. Provide high performance compute environment to enable variant calling pipeline to be executed as part of Station X data analysis workflow as they “curate” the TCGA data.
3. Additional use case is, for example, extracting a specific region from a BAM file across 200 different samples (ie 200 different BAM files). Evaluating GTFuse for this capability is of interest, but it likely of secondary concern at this time.

**The project**

**Project leader contact information:**

Name Sandeep Sanga, PhD

Phone (415) 944-4974

Email sandeep@stationxin.com

**Timescale and expected completion?**

Day 0 07/29/2013

Start Date 07/31/2013

End Date 08/14/2013

**Any there any major technical hurdles or points of interest identified?**

* Need to make sure Station X can have full access to VM.
* Is the local copy of Java up to date?
* Would prefer to load his image from AWS, understands this isn’t possible today.

**How will the project be judged a success – what are the key success criteria?**

Customer agrees to report back to Annai Systems within 5 days of completion of POC on the following:

1. Compute farm performance and speed
2. Ease of use of entire compute environment, (including GTFuse – nice to have)
3. GTFuse suitability for intended use (nice to have)
4. Any other performance or user experience aspects
5. Time to download and time to Job
6. Price per job compared to AWS

**How much budget been established for this and follow-on projects?**

2 x 8 core EC2

3 TB of Storage each

Associated S3 storage buckets

**What is the decision process?**

Sandeep will complete testing, then present findings and recommendation to Richard Gould for approval

**Project plan and weekly project reviews scheduled?**

* Day 0:
  + POC user account set up and training
* Review test plan and milestones with Annai and Customer teams
* Day 1:
* Support call with Annai to address any day 1 issues
* Customer to detail the relevant test plan and milestones
* Day 5:

Sandeep provided some great feedback.

* One of first test was to reproduce previous results.
* He believes he is getting better quality results.
  + Testing tumor vs. normal variant call.
  + Current solution:
    - Uses GeneTorrent and loads files into S3 bucket.
    - Going back a month later to run job.
    - Moves files into EC2 environment for analysis
    - AWS is 2 x 8 core instances
    - Believes it is the same memory config as BCF

Results:

Positive

* + Can log in.  Load tools etc.  Very similar to AWS.
  + Looking head to head is seeing better coverage using BCF.
  + Getting slightly different results.
    - Believes this is caused by data corruption through multiple transfers

Negative

* + Getting jobs completed much faster in AWS
    - BCF getting 26 or 27 Tumor normal pairs over 3 days
    - AWS getting 17 and 21 tumor normal pairs a day
  + AWS pricing for instance is getting pretty comparable.

Next steps

* + Sandeep to load performance monitoring tools to isolate the performance issue.
    - He believes it is likely a tuning issue, not a BCF issue.
    - He will provide feedback at day 11 call
* Day 11:
* Pre-completion call

**Current Environment**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Current AWS Solution ( x 2)** | | | | | | | | |
| **Instance Family** | **Instance Type** | **Processor Arch** | **vCPU** | **ECU** | **Memory (GiB)** | **Instance Storage (GB)** | **EBS-optimized Available** | **Network Performance** |
| Memory optimized | m2.4xlarge | 64-bit | 8 | 26 | 68.4 | 2 x 840 | Yes | High |
|  |  |  |  |  |  |  |  |  |
| **Current BCF Solution (x 1)** | | | | | | | | |
| **Instance Family** | **Instance Type** | **Processor Arch** | **vCPU** | **ECU** | **Memory (GiB)** | **Instance Storage (GB)** | **EBS-optimized Available** | **Network Performance** |
|  |  |  | 16 |  |  |  |  |  |

**AWS –**

* Station X
* downloads TCGA data from CGHub into an S3 bucket.
* Loads scheduler and pipeline tool into EC2.
* EC2 VMs pull a copy of the data from S3 into local storage for analysis.
* EC2 VMs push results back into S3

**BCF**

* Station X
* Loads scheduler and pipeline tool on BCF VM
* Pulls a copy of TCGA data from CGHub into local storage
* BCF pushes results to S3

**Sandeep ran a cost comparison and believes that we are significantly more expensive.**

* + - Their total project is:
    - Analyzing 9000 Exome pairs
    - Using BCF, they are currently averaging 11.22 pairs a day using
    - This will take 2.2 years to complete the analysis of all Exome pairs
  + Using AWS, they are currently averaging 21 pairs a day
    - 1.17 years to complete
      * Sandeep’s cost comparison
      * BCF total cost = $ 37,022.92
      * AWS total cost = $ 11,491
      * $ 1,000 Reservation Fee
      * $.51 Reserved price per machine hour (.51\*24\*365\*2)
      * “IF” BCF could match AWS performance, the total cost would be $ 19,784
      * This is $ 8,293 more than AWS
      * Problems with this analysis to address
      * Is this the correct “Up Front Reservation Fee”?
      * What about storage cost?
      * What about transfer cost?

**Performance issues**

* + - * Sandeep loaded NMON on VM to look for performance issues
      * Appears he is using the same number of Cores for BCF as he is for AWS
      * 8 for Compute
      * 8 for File Transfer
      * RAM is working at 100%
      * There is no easily identifiable / apparent reason for the performance difference.

**Next steps**

* + - Scott to address pricing issues and create an updated comparative cost analysis for Station X.
    - Mike to reach out to Station X for more details to identify the root cause of the performance discrepancy.
      * Sandeep responded to Mike’s request for more information with:
      * “What kind of performance data are you looking for? As you know, I'm primarily benching against throughput, but I monitor things like processor, memory, and disk i/o usage with a tool like nmon. As far as cloning the VM, please bear in mind that I am doing things that require individually granted access to the TCGA data (BAM-level files). Cloning my VM would breach that policy, wouldn't it, since the code, access-controlled files and my cghub key are on it?”
      * Mike to work with Matt and Chris to work up a benchmark environment within BCF and AWS. This is to address Station X’s issue as well as provide us with results we can provide new customer prospects.
    - Trial to be extended for 1 week to try and prove comparable performance with AWS.
    - Day 15 call still in place as the next touch point with customer.
* Day 15:
* Wrap up call