# DBU Estimating INSTRUCTIONS

Updated May 13, 2020

## Before We Begin…

* You can find the latest version of these instructions, or the Sheet itself, on Workramp.
* These instructions are for DBU Estimates, new workloads, “cold start.” If you’re trying to extrapolate out future DBUs based on past usage, that’s out of scope for this.

## Instructions

1. If you've gotten this far, presumably you have some new workloads that a customer wants to bring onto Databricks, and you've been tasked with estimating the DBUs it (they) will consume. Congratulations! You can move onto Step 2 now!
2. The first thing you should do is update the metadata on the *“Opp Info”* tab. This includes some basic information like Customer Name; use-case names; deployment timelines; etc. NOTE: that the use-case names are particularly important as those flow through the rest of the tabs in the Sheet.
3. Next, be sure to customize the *“Architecture”* tab with a picture that’s more representative for your particular customer environment.
4. On the *"Use Case Job Details (CUSTOMER VERSION)"* tab, start collecting the relevant data points about that workload. This will include at minimum:  
   1. Types of Jobs (Batch, Streaming. Data Science, etc.)
   2. Number of jobs
   3. Job complexity / data sizes
   4. Time to run the job in the existing environment (e.g. Ideally based on POC run-times and comparison to POC workfload or a fraction of existing platforms run-times, new SLAs required, etc.)
   5. Number of Interactive/Analytics clusters needed (based on # of business units / Teams, and # of concurrent users)
   6. Number of Environments needed
5. Continue filling out the information for the other use-cases, as needed, until you’ve done what you can and you need to involve the customer directly.
6. Right-click on the *"Use Case Job Details (CUSTOMER VERSION)"* tab and select "Copy to -> New Spreadsheet." This will take the information and make an entirely new Google Sheet which you can share directly with your customer.  
     
   The difference between this tab and the *“Use Case Job Details”* tab is that the latter includes pricing information whereas the customer-facing version doesn’t. This should be the only difference between those two tabs.
7. Share the new document with your customer, and work with them to help fill in the missing information.
   1. Work with your customer to validate assumptions. These template estimates are provided with the following assumptions built-in:  
      1. 100% of usage is predicted for Production environment; 10-25% of that for Dev; and an additional 10-25% for QA environment
      2. Ad-hoc clusters estimated for 22 days of activity per month, with 8 hrs of uptime
      3. 3-5% monthly data growth
      4. Estimated ETL workloads for automated jobs
      5. Ad-hoc clusters used by Data Science, Developer, QA / Support, and Prod Support Teams
   2. This might be the most important step - work WITH your customer (typically the Technical Architects or Lead Data Scientist) to ensure they know how the data being collected will impact the estimates. Your customer needs to own these estimates as if they built it themselves. They know their environment better than we do, so ensure you're engaging with the right people to get their perspective.
8. At this point, you'll likely need to benchmark or profile your workloads in Databricks. A benchmarking exercise (which could be a full-fledged POC) will likely be done at this point, providing even more data points and help you further refine your estimates. Fill that data in here as needed.
9. Once you're confident you have the information you need to finalize your estimates, you'll need to work with your AE to finalize your DBU estimates.  
     
   Go to the *"Summary: DBU Ramp-up"* tab to take your monthly estimates and make them a bit more real-world for your specific customer environment. This methodology provides the ability to "ramp-up" DBus associated with each use-case. For example, the second use-case you scoped out might be dependent on the first one being implemented. If that first use-case takes 2 months to implement, you might want to start counting DBUs for the second use-case starting in month 3.  
     
   For example, sometimes the customer wants to roll out the project by stores (e.g. CVS) or machinery in the case of IoT (e.g. Komatsu) or factories which result in gradual ramp-up of the DBUs over time.
10. Review the estimates with your AE and work to get them translated into a Demand Capacity Sheet. There are a few versions of this being used by the Sales Team today, but your AE (or Sales Leader) will be able to point you to the right version for your region. The one being used in Enterprise West is in the main Google Drive folder linked to above.  
      
    In most situations, you’ll be taking the ramped-up annual numbers and dividing by 12 to get an “average monthly DBU” number. This information is in the *“Summary: DBU Ramp-up”* tab in Column U.

## Special Note for Streaming DBUs

For bigger/high-frequency mission critical streams (e.g. 100K - 1 millions messages/sec) use separate streaming clusters per stream (SinglePlex). But if you plan to run multiple small-medium streams per cluster (Multiplexing), ideally don’t go over 30-40 streams on the same cluster. ( E.g. Not Streaming, but trickling real-time CDC from hundreds or thousands of source tables).