# Assignment

Within one of our brands we have a product called VPS. VPS Stands for Virtual Private Server and is a full server as one would expect, just not a physical one. We are curious to see whether we can predict the churn based upon the product usage.

You have 2 files: *vps\_churn\_data.csv* and *vps\_test\_data.csv* . The first one of these files contain the usages of VPSses and a column called *is\_churn* which specifies whether that VPS has churned at that period. The second file contains new data for VPSses, with the *is\_churn­* column left blank. It’s your job to predict wether this VPS will churn or not.

So, you need to fill the is\_churn column in the *vps\_test\_data.csv* file. This data needs to eventually reach the stakeholders so they can do something with this. Please deliver the data in such a way the stakeholders an directly use it, or in such a way the data analysts can actually use it to create insights & reporting for it.

We would like to ask you to use python or R for this, preferred in combination with Spark. The time window for this is **48 hours** from now on.

## Some background info

* The data is already a downsampled set of the actual full imbalanced dataset
* We measure different aspects of a VPS: CPU load, disk octets read & write, disk iops read & write, network tx & rx. At least some are relevant, but it could be that some aren’t.
* Which models etc. to choose is fully up to you!
* How you deliver the insights, is also up to you, just make sure the people “further down the data pipeline” can use it.