

Hard Disk Drive SMART Data Analysis using Apache Spark

ECE 590 – Big Data Technologies-Spring 2021

Team – 18

Shamili M Tetali
George Mason University
Fairfax, Virginia, 22030
stetali@gmu.edu

Matiyas Maru
George Mason University
Fairfax, Virginia, 22030
Mmaru@gmu.edu

Muhammed Hassan
George Mason University
Fairfax, Virginia, 22030
Mhassa9@gmu.edu

Abstract

The enormous growth of data acquisition use cases in recent years has indeed demanded more storage resources, primarily hard disk drives (HDD) where data resides. Based on the organization's IT design choices these physical storage components could either be housed at Cloud storage or on-premises datacenters. Data availability from any of these storage implementations heavily depends on reliability characteristics of the underlying infrastructure. Apparently, of all the hardware components, hard disks experience high failure rates and in several cases its quite uncertain to accurately predict failures ahead of the impact, thereby losing service credibility and negatively influences financial prospects.

The scope of this project is primarily based on selected research papers[2][3][4] and data collected from Backblaze datasets. In this paper, we will leverage Big Data analytic concepts and characterize data, based on Hard Disks inherent feature called Self-Monitoring Analysis and Reporting Technology (SMART) attributes. Essentially, for the purpose of this study we obtained historical datasets from Backblaze[1] and considered SMART parameters that indicate abnormal hard disk behavior. Data analysis is performed by using Apache Spark implementation and finally established an analytical framework that will help to understand and determine co-relation between parameters (like temperature, capacity, manufacturer etc).

I. Introduction.

As demand for IT Infrastructure intensifies, deployments surged at an unprecedented rate and massively expanding at global scale. Reliability and availability being important factors, service providers and engineers focuses their attention to maintaining the ecosystem and aims to delivering uninterrupted service. Therefore, fault prevention requires advanced

data analysis methods to early detection of failures analysis. Proactive detection and identification of abnormalities is the key to meeting the standards. Disks are among the most frequently failed components and observing abnormalities to predicting the impending failure of hard disks in the field can help systems at large organizations, datacenters and other crucial places to take corrective actions before the failure to avoid loss of data and performance degradation.

Data reliability, availability and low ownership costs are basic expectations from data users and it's essential for engineers to understanding the nature of failures and come up with predictive analysis mechanisms. With major advancements in machine learning, data mining etc, it has become possible to devise data analytic methods to proactively determine disk failure rates and preemptively solve large scale manufacturing concerns, thereby reducing the TCO and costs associated to HDD manufacturing and return merchandize processes.

The primary goal of this study is to implement Big Data analytics on HDD Smart attribute datasets and to understanding the relationship between various parameters that could negatively impact disk drive performance and premature failures. A significant size of 52 million records are considered towards the scope of this project and utilize Apache Spark as the analytics engine. The study will also demonstrate Big Data analytic Framework capabilities for various reporting and how we could derive value added solutions for Hard Disk Drive predictive failures.

II. Data Flow Architecture.

Fig(1) shows the schematic diagram of our analysis using Apache Spark. The data is collected from Backblaze, stored in our Local Machine. These are further analyzed using Apache Zeppelin notebook

which triggers the spark jobs computation and visualize the results.

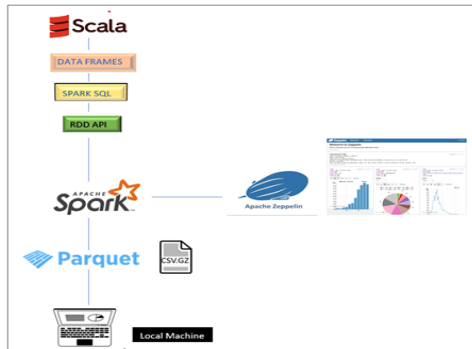


Fig (1) : Data Flow Architecture

For Analysis visualization we used open source software as listed below.

1. Apache Spark computation engine: version-2.4.7
2. Apache Zeppelin web notebook for querying and visualization: version-0.8.2
3. SCALA programming language : version-2.11.12

III. Implementation.

Data Collection:

The source of Data is collected from Backblaze which is primarily a Cloud Storage and Backup company[link1]. hard drive dataset. For the purpose of this project, we reviewed data files for year 2020 (Q1-Q4) that consisted 52286398 records (approx. 52 million).

Further breakdown of Data files specifics:

Total Files Size	16.37GB
Files Count	370 files
Drive Count	162,299
Drive Failures	1,302
Drive Days	51.2M
Drive Capacity Range	240GB – 18TB

The first row of each file contains the column names, the remaining rows are the actual data. The columns are as follows:

- Date** – The date of the file in yyyy-mm-dd format.
- Serial Number** – Mfg. assigned serial number of the drive.
- Model** – Mfg. assigned model number of the drive.
- Capacity** – The drive capacity in bytes.

Failure – Contains a “0” if the drive is OK. Contains a “1” if this is the last day the drive was operational before failing.

The remaining columns are Smart Attributes associated to Normalized and RAW values that ranges from 1 – 255 and each value signify a drive operating characteristic as reported by the drive built in smart function.

Hard Disk SMART Data Analysis:

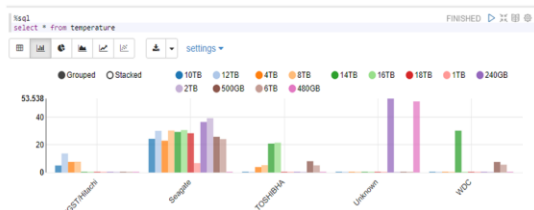
The data consists of daily snapshots of the SMART statistics and a failure label for all operational hard drives in a data center in 2020. SMART stats are meant to be indicators of drive reliability and should, in theory, provide good input features to a predictive model of drive failure. The first step is to get all data in csv format arrangement and combine each month data and further archived those files for efficiency purposes. Data is converted to parquet format which allows compression schemes to be specified on a per-column basis. The data is then read via Spark and preview of most used column data is shown below in table[1].

Datafiles in consists of both “normalized” and “raw” columns for each of the SMART attribute. “Normalized” columns are ignored for the purpose of this analysis as standardized RAW data would suffice the need and provide detailed information required to determining drive statistics. Manufacturer-specific normalizations would be applicable to partial datasets due to transformations and therefore creates discrepancies. Hence, for the scope of this analysis we considered denormalized Raw values to establishing a model across all manufacturers drives.

		serial	model	capacity	failure	smart_attr_1	smart_attr_2	smart_attr_3	smart_attr_4	smart_attr_5	smart_attr_6	smart_attr_7	smart_attr_8	smart_attr_9	smart_attr_10	smart_attr_11	smart_attr_12	smart_attr_13	smart_attr_14	smart_attr_15	smart_attr_16	smart_attr_17	smart_attr_18	smart_attr_19	smart_attr_20	smart_attr_21	smart_attr_22	smart_attr_23	smart_attr_24	smart_attr_25	smart_attr_26	smart_attr_27	smart_attr_28	smart_attr_29	smart_attr_30	smart_attr_31	smart_attr_32	smart_attr_33	smart_attr_34	smart_attr_35	smart_attr_36	smart_attr_37	smart_attr_38	smart_attr_39	smart_attr_40	smart_attr_41	smart_attr_42	smart_attr_43	smart_attr_44	smart_attr_45	smart_attr_46	smart_attr_47	smart_attr_48	smart_attr_49	smart_attr_50	smart_attr_51	smart_attr_52	smart_attr_53	smart_attr_54	smart_attr_55	smart_attr_56	smart_attr_57	smart_attr_58	smart_attr_59	smart_attr_60	smart_attr_61	smart_attr_62	smart_attr_63	smart_attr_64	smart_attr_65	smart_attr_66	smart_attr_67	smart_attr_68	smart_attr_69	smart_attr_70	smart_attr_71	smart_attr_72	smart_attr_73	smart_attr_74	smart_attr_75	smart_attr_76	smart_attr_77	smart_attr_78	smart_attr_79	smart_attr_80	smart_attr_81	smart_attr_82	smart_attr_83	smart_attr_84	smart_attr_85	smart_attr_86	smart_attr_87	smart_attr_88	smart_attr_89	smart_attr_90	smart_attr_91	smart_attr_92	smart_attr_93	smart_attr_94	smart_attr_95	smart_attr_96	smart_attr_97	smart_attr_98	smart_attr_99	smart_attr_100	smart_attr_101	smart_attr_102	smart_attr_103	smart_attr_104	smart_attr_105	smart_attr_106	smart_attr_107	smart_attr_108	smart_attr_109	smart_attr_110	smart_attr_111	smart_attr_112	smart_attr_113	smart_attr_114	smart_attr_115	smart_attr_116	smart_attr_117	smart_attr_118	smart_attr_119	smart_attr_120	smart_attr_121	smart_attr_122	smart_attr_123	smart_attr_124	smart_attr_125	smart_attr_126	smart_attr_127	smart_attr_128	smart_attr_129	smart_attr_130	smart_attr_131	smart_attr_132	smart_attr_133	smart_attr_134	smart_attr_135	smart_attr_136	smart_attr_137	smart_attr_138	smart_attr_139	smart_attr_140	smart_attr_141	smart_attr_142	smart_attr_143	smart_attr_144	smart_attr_145	smart_attr_146	smart_attr_147	smart_attr_148	smart_attr_149	smart_attr_150	smart_attr_151	smart_attr_152	smart_attr_153	smart_attr_154	smart_attr_155	smart_attr_156	smart_attr_157	smart_attr_158	smart_attr_159	smart_attr_160	smart_attr_161	smart_attr_162	smart_attr_163	smart_attr_164	smart_attr_165	smart_attr_166	smart_attr_167	smart_attr_168	smart_attr_169	smart_attr_170	smart_attr_171	smart_attr_172	smart_attr_173	smart_attr_174	smart_attr_175	smart_attr_176	smart_attr_177	smart_attr_178	smart_attr_179	smart_attr_180	smart_attr_181	smart_attr_182	smart_attr_183	smart_attr_184	smart_attr_185	smart_attr_186	smart_attr_187	smart_attr_188	smart_attr_189	smart_attr_190	smart_attr_191	smart_attr_192	smart_attr_193	smart_attr_194	smart_attr_195	smart_attr_196	smart_attr_197	smart_attr_198	smart_attr_199	smart_attr_200	smart_attr_201	smart_attr_202	smart_attr_203	smart_attr_204	smart_attr_205	smart_attr_206	smart_attr_207	smart_attr_208	smart_attr_209	smart_attr_210	smart_attr_211	smart_attr_212	smart_attr_213	smart_attr_214	smart_attr_215	smart_attr_216	smart_attr_217	smart_attr_218	smart_attr_219	smart_attr_220	smart_attr_221	smart_attr_222	smart_attr_223	smart_attr_224	smart_attr_225	smart_attr_226	smart_attr_227	smart_attr_228	smart_attr_229	smart_attr_230	smart_attr_231	smart_attr_232	smart_attr_233	smart_attr_234	smart_attr_235	smart_attr_236	smart_attr_237	smart_attr_238	smart_attr_239	smart_attr_240	smart_attr_241	smart_attr_242	smart_attr_243	smart_attr_244	smart_attr_245	smart_attr_246	smart_attr_247	smart_attr_248	smart_attr_249	smart_attr_250	smart_attr_251	smart_attr_252	smart_attr_253	smart_attr_254	smart_attr_255	smart_attr_256	smart_attr_257	smart_attr_258	smart_attr_259	smart_attr_260	smart_attr_261	smart_attr_262	smart_attr_263	smart_attr_264	smart_attr_265	smart_attr_266	smart_attr_267	smart_attr_268	smart_attr_269	smart_attr_270	smart_attr_271	smart_attr_272	smart_attr_273	smart_attr_274	smart_attr_275	smart_attr_276	smart_attr_277	smart_attr_278	smart_attr_279	smart_attr_280	smart_attr_281	smart_attr_282	smart_attr_283	smart_attr_284	smart_attr_285	smart_attr_286	smart_attr_287	smart_attr_288	smart_attr_289	smart_attr_290	smart_attr_291	smart_attr_292	smart_attr_293	smart_attr_294	smart_attr_295	smart_attr_296	smart_attr_297	smart_attr_298	smart_attr_299	smart_attr_300	smart_attr_301	smart_attr_302	smart_attr_303	smart_attr_304	smart_attr_305	smart_attr_306	smart_attr_307	smart_attr_308	smart_attr_309	smart_attr_310	smart_attr_311	smart_attr_312	smart_attr_313	smart_attr_314	smart_attr_315	smart_attr_316	smart_attr_317	smart_attr_318	smart_attr_319	smart_attr_320	smart_attr_321	smart_attr_322	smart_attr_323	smart_attr_324	smart_attr_325	smart_attr_326	smart_attr_327	smart_attr_328	smart_attr_329	smart_attr_330	smart_attr_331	smart_attr_332	smart_attr_333	smart_attr_334	smart_attr_335	smart_attr_336	smart_attr_337	smart_attr_338	smart_attr_339	smart_attr_340	smart_attr_341	smart_attr_342	smart_attr_343	smart_attr_344	smart_attr_345	smart_attr_346	smart_attr_347	smart_attr_348	smart_attr_349	smart_attr_350	smart_attr_351	smart_attr_352	smart_attr_353	smart_attr_354	smart_attr_355	smart_attr_356	smart_attr_357	smart_attr_358	smart_attr_359	smart_attr_360	smart_attr_361	smart_attr_362	smart_attr_363	smart_attr_364	smart_attr_365	smart_attr_366	smart_attr_367	smart_attr_368	smart_attr_369	smart_attr_370	smart_attr_371	smart_attr_372	smart_attr_373	smart_attr_374	smart_attr_375	smart_attr_376	smart_attr_377	smart_attr_378	smart_attr_379	smart_attr_380	smart_attr_381	smart_attr_382	smart_attr_383	smart_attr_384	smart_attr_385	smart_attr_386	smart_attr_387	smart_attr_388	smart_attr_389	smart_attr_390	smart_attr_391	smart_attr_392	smart_attr_393	smart_attr_394	smart_attr_395	smart_attr_396	smart_attr_397	smart_attr_398	smart_attr_399	smart_attr_400	smart_attr_401	smart_attr_402	smart_attr_403	smart_attr_404	smart_attr_405	smart_attr_406	smart_attr_407	smart_attr_408	smart_attr_409	smart_attr_410	smart_attr_411	smart_attr_412	smart_attr_413	smart_attr_414	smart_attr_415	smart_attr_416	smart_attr_417	smart_attr_418	smart_attr_419	smart_attr_420	smart_attr_421	smart_attr_422	smart_attr_423	smart_attr_424	smart_attr_425	smart_attr_426	smart_attr_427	smart_attr_428	smart_attr_429	smart_attr_430	smart_attr_431	smart_attr_432	smart_attr_433	smart_attr_434	smart_attr_435	smart_attr_436	smart_attr_437	smart_attr_438	smart_attr_439	smart_attr_440	smart_attr_441	smart_attr_442	smart_attr_443	smart_attr_444	smart_attr_445	smart_attr_446	smart_attr_447	smart_attr_448	smart_attr_449	smart_attr_450	smart_attr_451	smart_attr_452	smart_attr_453	smart_attr_454	smart_attr_455	smart_attr_456	smart_attr_457	smart_attr_458	smart_attr_459	smart_attr_460	smart_attr_461	smart_attr_462	smart_attr_463	smart_attr_464	smart_attr_465	smart_attr_466	smart_attr_467	smart_attr_468	smart_attr_469	smart_attr_470	smart_attr_471	smart_attr_472	smart_attr_473	smart_attr_474	smart_attr_475	smart_attr_476	smart_attr_477	smart_attr_478	smart_attr_479	smart_attr_480	smart_attr_481	smart_attr_482	smart_attr_483	smart_attr_484	smart_attr_485	smart_attr_486	smart_attr_487	smart_attr_488	smart_attr_489	smart_attr_490	smart_attr_491	smart_attr_492	smart_attr_493	smart_attr_494	smart_attr_495	smart_attr_496	smart_attr_497	smart_attr_498	smart_attr_499	smart_attr_500	smart_attr_501	smart_attr_502	smart_attr_503	smart_attr_504	smart_attr_505	smart_attr_506	smart_attr_507	smart_attr_508	smart_attr_509	smart_attr_510	smart_attr_511	smart_attr_512	smart_attr_513	smart_attr_514	smart_attr_515	smart_attr_516	smart_attr_517	smart_attr_518	smart_attr_519	smart_attr_520	smart_attr_521	smart_attr_522	smart_attr_523	smart_attr_524	smart_attr_525	smart_attr_526	smart_attr_527	smart_attr_528	smart_attr_529	smart_attr_530	smart_attr_531	smart_attr_532	smart_attr_533	smart_attr_534	smart_attr_535	smart_attr_536	smart_attr_537	smart_attr_538	smart_attr_539	smart_attr_540	smart_attr_541	smart_attr_542	smart_attr_543	smart_attr_544	smart_attr_545	smart_attr_546	smart_attr_547	smart_attr_548	smart_attr_549	smart_attr_550	smart_attr_551	smart_attr_552	smart_attr_553	smart_attr_554	smart_attr_555	smart_attr_556	smart_attr_557	smart_attr_558	smart_attr_559	smart_attr_560	smart_attr_561	smart_attr_562	smart_attr_563	smart_attr_564	smart_attr_565	smart_attr_566	smart_attr_567	smart_attr_568	smart_attr_569	smart_attr_570	smart_attr_571	smart_attr_572	smart_attr_573	smart_attr_574	smart_attr_575	smart_attr_576	smart_attr_577	smart_attr_578	smart_attr_579	smart_attr_580	smart_attr_581	smart_attr_582	smart_attr_583	smart_attr_584	smart_attr_585	smart_attr_586	smart_attr_587	smart_attr_588	smart_attr_589	smart_attr_590	smart_attr_591	smart_attr_592	smart_attr_593	smart_attr_594	smart_attr_595	smart_attr_596	smart_attr_597	smart_attr_598	smart_attr_599	smart_attr_600	smart_attr_601	smart_attr_602	smart_attr_603	smart_attr_604	smart_attr_605	smart_attr_606	smart_attr_607	smart_attr_608	smart_attr_609	smart_attr_610	smart_attr_611	smart_attr_612	smart_attr_613	smart_attr_614	smart_attr_615	smart_attr_616	smart_attr_617	smart_attr_618	smart_attr_619	smart_attr_620	smart_attr_621	smart_attr_622	smart_attr_623	smart_attr_624	smart_attr_625	smart_attr_626	smart_attr_627	smart_attr_628	smart_attr_629	smart_attr_630	smart_attr_631	smart_attr_632	smart_attr_633	smart_attr_634	smart_attr_635	smart_attr_636	smart_attr_637	smart_attr_638	smart_attr_639	smart_attr_640	smart_attr_641	smart_attr_642	smart_attr_643	smart_attr_644	smart_attr_645	smart_attr_646	smart_attr_647	smart_attr_648	smart_attr_649	smart_attr_650	smart_attr_651	smart_attr_652	smart_attr_653	smart_attr_654	smart_attr_655	smart_attr_656	smart_attr_657	smart_attr_658	smart_attr_659	smart_attr_660	smart_attr_661	smart_attr_662	smart_attr_663	smart_attr_664	smart_attr_665	smart_attr_666	smart_attr_667	smart_attr_668	smart_attr_669	smart_attr_670	smart_attr_671	smart_attr_672	smart_attr_673	smart_attr_674	smart_attr_675	smart_attr_676	smart_attr_677	smart_attr_678	smart_attr_679	smart_attr_680	smart_attr_681	smart_attr_682	smart_attr_683	smart_attr_684	smart_attr_685	smart_attr_686	smart_attr_687	smart_attr_688	smart_attr_689	smart_attr_690	smart_attr_691	smart_attr_692	smart_attr_693	smart_attr_694	smart_attr_695	smart_attr_696	smart_attr_697	smart_attr_698	smart_attr_699	smart_attr_700	smart_attr_701	smart_attr_702	smart_attr_703	smart_attr_704	smart_attr_705	smart_attr_706	smart_attr_707	smart_attr_708	smart_attr_709	smart_attr_710	smart_attr_711	smart_attr_712	smart_attr_713	smart_attr_714	smart_attr_715	smart_attr_716	smart_attr_717	smart_attr_718	smart_attr_719	smart_attr_720	smart_attr_721	smart_attr_722	smart_attr_723	smart_attr_724	smart_attr_725	smart_attr_726	smart_attr_727	smart_attr_728	smart_attr_729	smart_attr_730	smart_attr_731	smart_attr_732	smart_attr_733	smart_attr_734	smart_attr_735	smart_attr_736	smart_attr_737	smart_attr_738	smart_attr_739	smart_attr_740	smart_attr_741	smart_attr_742	smart_attr_743	smart_attr_744	smart_attr_745	smart_attr_746	smart_attr_747	smart_attr_748	smart_attr_749	smart_attr_750	smart_attr_751	smart_attr_752	smart_attr_753	smart_attr_754	smart_attr_755	smart_attr_756	smart_attr_757	smart_attr_758	smart_attr_759	smart_attr_760	smart_attr_761	smart_attr_762	smart_attr_763	smart_attr_764	smart_attr_765	smart_attr_766	smart_attr_767	smart_attr_768	smart_attr_769	smart_attr_770	smart_attr_771	smart_attr_772	smart_attr_773	smart_attr_774	smart_attr_775	smart_attr_776	smart_attr_777	smart_attr_778	smart_attr_779	smart_attr_780	smart_attr_781	smart_attr_782	smart_attr_783	smart_attr_784	smart_attr_785	smart_attr_786	smart_attr_787	smart_attr_788	smart_attr_789	smart_attr_790	smart_attr_791	smart_attr_792	smart_attr_793	smart_attr_794	smart_attr_795	smart_attr_796	smart_attr_797	smart_attr_798	smart_attr_799	smart_attr_800	smart_attr_801	smart_attr_802	smart_attr_803	smart_attr_804	smart_attr_805	smart_attr_806	smart_attr_807	smart_attr_808	smart_attr_809	smart_attr_810	smart_attr_811	smart_attr_812	smart_attr_813	smart_attr_814	smart_attr_815	smart_attr_816	smart_attr_817	smart_attr_818	smart_attr_819	smart_attr_820	smart_attr_821	smart_attr_822	smart_attr_823	smart_attr_824	smart_attr_825	smart_attr_826	smart_attr_827	smart_attr_828	smart_attr_829	smart_attr_830	smart_attr_831	smart_attr_832	smart_attr_833	smart_attr_834	smart_attr_835	smart_attr_836	smart_attr_837	smart_attr_838	smart_attr_839	smart_attr_840	smart_attr_841	smart_attr_842	smart_attr_843	smart_attr_844	smart_attr_845	smart_attr_846	smart_attr_847	smart_attr_848	smart_attr_849	smart_attr_850	smart_attr_851	smart_attr_852	smart_attr_853	smart_attr_854	smart_attr_855	smart_attr_856	smart_attr_857	smart_attr_858	smart_attr_859	smart_attr_860	smart_attr_861	smart_attr_862	smart_attr_863	smart_attr_864	smart_attr_865	smart_attr_866	smart_attr_867	smart_attr_868	smart_attr_869	smart_attr_870	smart_attr_871	smart_attr_872	smart_attr_873	smart_attr_874	smart_attr_875	smart_attr_876	smart_attr_877	smart_attr_878	smart_attr_879	smart_attr_880	smart_attr_881	smart_attr_882	smart_attr_883	smart_attr_884	smart_attr_885	smart_attr_886	smart_attr_887	smart_attr_888	smart_attr_889	smart_attr_890	smart_attr_891	smart_attr_892	smart_attr_893	smart_attr_894	smart_attr_895	smart_attr_896	smart_attr_897	smart_attr_898	smart_attr_899	smart_attr_900	smart_attr_901	smart_attr_902	smart_attr_903	smart_attr_904	smart_attr_905	smart_attr_906	smart_attr_907	smart_attr_908	smart_attr_909	smart_attr_910	smart_attr_911	smart_attr_912	smart_attr_913	smart_attr_914	smart_attr_915	smart_attr_916	smart_attr_917	smart_attr_918	smart_attr_919	smart_attr_920	smart_attr_921	smart_attr_922	smart_attr_923	smart_attr_924	smart_attr_925	smart_attr_926	smart_attr_927	smart_attr_928	smart_attr_929	smart_attr_930	smart_attr_931	smart_attr_932	smart_attr_933	smart_attr_934	smart_attr_935	smart_attr_936	smart_attr_937	smart_attr_938	smart_attr_939	smart_attr_940	smart_attr_941	smart_attr_942	smart_attr_943	smart_attr_944	smart_attr_945	smart_attr_946	smart_attr_947	smart_attr_948	smart_attr_949	smart_attr_950	smart_attr_951	smart_attr_952	smart_attr_953	smart_attr_954	smart_attr_955	smart_attr_956	smart_attr_957	smart_attr_958	smart_attr_959	smart_attr_960	smart_attr_961	smart_attr_962	smart_attr_963	smart_attr_964	smart_attr_965	smart_attr_966	smart_attr_967	smart_attr_968	smart_attr_969	smart_attr_970	smart_attr_971	smart_attr_972	smart_attr_973	smart_attr_974	smart_attr_975	smart_attr_976	smart_attr_977	smart_attr_978	smart_attr_979	smart_attr_980	smart_attr_981	smart_attr_982	smart_attr_983	smart_attr_984	smart_attr_985	smart_attr_986	smart_attr_987	smart_attr_988	smart_attr_989	smart_attr_990	smart_attr_991	smart_attr_992	smart_attr_993	smart_attr_994	smart_attr_995	smart_attr_996	smart_attr_997	smart_attr_998	smart_attr_999	smart_attr_1000
2020-12-01	000																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																												

Analysis of Temperature on hard drives:

We analyzed the Smart_194_raw value which represent the temperature of the drive for various manufacturers. Through the visualization below, the distribution of drive temperatures for our four most popular drives. Depicted in bar graph below Fig(2), all drives were well operated between 0° (or 5°) to 60° as specified by manufacturers and within threshold. However, typical operating temperature range of Hard Disk drives could potentially vary by specific disk design characteristics.



Fig(2) : Bar chart on MFG vs TEMP(avg)

Computing Annual Failure Rate:

Considering a given group of drives (i.e. model, manufacturer, etc.) an attempt is made to compute the AFR for a period of observation as follows:

$$AFR = (\text{Drive Failures} / (\text{Drive Days} / 366)) * 100$$

where:

Drive Failures are the number of drives that failed during the period of observation.

Drive Days is number of days all of the observed drives were operational during the period of observation.

There are 366 days in 2020, obviously in non-leap years we would consider 365.

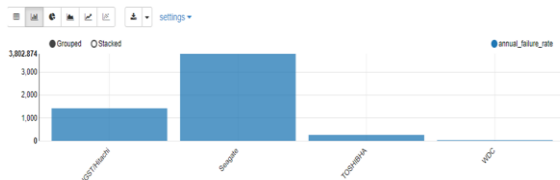


Fig (3) : Annual Failure Rate 2020 stats

manufacturer	model	capacity_bytes	drivedays	failures	temperature	annual_failure_rate
HGST/Hitachi	HGST HD55C4040ALE630	4TB	9276	1	0	3.945666
HGST/Hitachi	HGST HM55C4040ALE640	4TB	1083641	8	19.5	0.2702
HGST/Hitachi	HGST HM55C4040BLE640	4TB	4662611	34	200.1470588	0.266889
HGST/Hitachi	HGST HU721212ALE600	12TB	820272	7	11.42857143	0.312335
HGST/Hitachi	HGST HU721212ALE604	12TB	275036	9	24.11111111	1.197661
HGST/Hitachi	HGST HU721212ALN604	12TB	3968303	50	3.84	0.461154
HGST/Hitachi	HGST HU728080ALE600	8TB	371930	3	13.33333333	0.295217
Seagate	ST1000NM0086	10TB	110451	7	28.57142857	2.319581
Seagate	ST12000NM0007	12TB	2314237	66	28.3030303	1.0438
Seagate	ST12000NM0008	12TB	1740779	49	32.85714286	1.030228
Seagate	ST12000NM001G	12TB	610091	12	33.83333333	0.719893
Seagate	ST14000NM001G	14TB	431057	13	32.53846154	1.103798
Seagate	ST18000NM000J	18TB	5491	2	30	13.330905
Seagate	ST4000DM000	4TB	1745899	83	23.79518072	1.739963
Seagate	ST500LM012 HN	500GB	170910	34	5.352941176	7.281025
Seagate	ST8000DM002	8TB	900304	28	33.28571429	1.138282
Seagate	ST8000NM0055	8TB	1325815	47	36.9787234	1.297466
Seagate	Seagate SSD	240GB	9902	1	38	3.696223
TOSHIBA	TOSHIBA MG07ACA14TA	14TB	4100116	102	12.56862745	0.910511
TOSHIBA	TOSHIBA MQ01ABF050	500GB	143891	96	7.625	24.418483
TOSHIBA	TOSHIBA MQ01ABF050M	500GB	145672	32	10.3125	8.03998
WDC	WDC WD5000LPCX	500GB	19476	1	0	1.879236
WDC	WDC WD5000LPVX	500GB	73535	10	0	4.977222
WDC	WDC WU721414ALE6L4	14TB	226848	1	40	0.161342

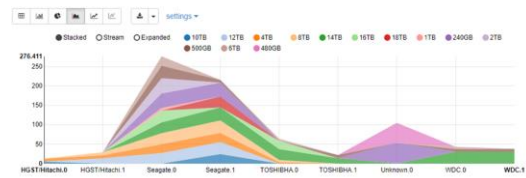
Table : Annual Failure Rate based on 2020 stats

After performing a comparative analysis, it's determined that on an average Seagate Drives shows highest rate of Annual failure and then followed by Hitachi as shown in Fig(3) and the table above.

IV. Results & Discussions.

Effect of Disk Temp on Predictive Failure Abnormalities:

Overall, there is no direct correlation between operating temperature and failure rates. However, from the observation, Seagate drives models from 2020 are generating more heat while experiencing predictive failure abnormalities and therefore disrupts thermal profiles in Datacenter Environments. This is an important metric as to understanding HDD behavior, which in this case Seagate generate elevated heat levels and contribute to thermal profile variations in storage deployments and thereby increasing Operating expenses. Refer Fig(4)



Fig(4) : Temperature vs Failure rate

Relationship between Predictive failure and 5 main SMART Parameters:

SMART readings presented by Hard Disks can be out-of-bounds, noisy, or inaccurate based on disk design and operating conditions, and therefore sometimes have quite a bit of missing data. One step to address these problems is to filter out columns where a lot of the entries are null.

The 5 Key SMART attributes for predictive errors:

SMART 5 - Reallocated_Sector_Count.
 SMART 187 - Reported_Uncorrectable_Errors.
 SMART 188 - Command_Timeout.
 SMART 197 - Current_Pending_Sector_Count.
 SMART 198 - Offline_Uncorrectable

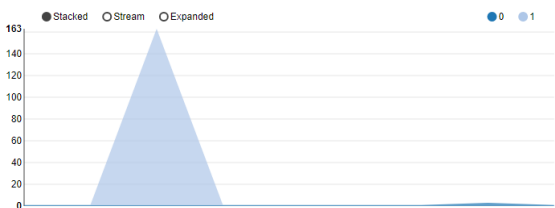
As stated by Backblaze, When RAW value for one of these five attributes is greater than zero, we have a clear indication to consider the value and investigate. Below graphs are to see whether there is a true relation between these SMART parameters and the failure.

Based on the observations, except for SMART 188, the definitions described above from Backblaze holds true for every other SMART value.

SMART 5



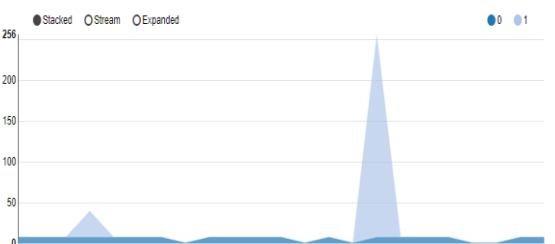
SMART 187



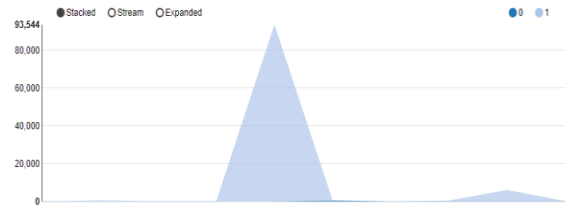
SMART 188



SMART 197



SMART 198



Considering the scope for next level analysis we have implemented **Linear Regression Model** with these smart parameters as features and failure column as label. However, the coefficients are sometimes visible as blank values and thus we cannot perform our analysis further to display predictions. Therefore, in the future scope we would like to investigate which regressions would best suit to predict the drive failure using machine learning training models.

V. Conclusion.

Proactive monitoring and management capabilities are the first steps towards improving current standards and modelling systems for gaining Infrastructure Insights. As summarized in each section, various methods and approaches are taken to demonstrate the powerful capabilities of Big Data concepts that included Spark and other supporting utilities. In this study, Hard Disks SMARTs were analyzed for measuring Hard Disk Drive Quality that essentially adds value to improving the availability and reliability of underlying storage. Primarily based upon engineering analysis, we considered the characteristics of each query, the type of data and size of data set were the main factors to optimally demonstrating the capabilities of Big Data Analytics and estimated the failure rates, therefore plotting information to minimizing the scope of impact to infrastructure. The methodologies described can provide proactive and valuable Insights to HDD Manufacturers and Customer Deployments. The paper also establishes relationship between various characteristics of the Hard Disks to understanding the operational overheads like thermals and discusses its impact.

VI. Future Scope.

In this paper, we proposed and established a solution that can be further extended to developing Telemetry & Transform Management system with advanced machine learning techniques. Applying deep data analytics to telemetry data to potentially enable Self Managing, Self-Healing and Self Optimizing features and improving the availability and performance of Storage infrastructure.

References:

1. <https://www.backblaze.com/b2/hard-drive-test-data.html>
2. D. D. Mishra, S. Pathan and C. Murthy, "Apache Spark Based Analytics of Squid Proxy Logs," 2018 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), 2018, pp. 1-6, doi: 10.1109/ANTS.2018.8710044.
3. Singh, Archana & Mittal, Mamta & Kapoor, Namita. (2018). Data Processing Framework Using Apache and Spark Technologies in Big Data. 10.1007/978-981-13-0550-4_5.
4. Botezatu, Mirela & Giurgiu, Ioana & Bogojeska, Jasmina & Wiesmann, Dorothea. (2016). Predicting Disk Replacement towards Reliable Data Centers. 39-48. 10.1145/2939672.2939699.
5. <https://renovacloud.com/an-introduction-to-and-evaluation-of-apache-spark-for-big-data-architectures/?lang=en>