

**PROJECT PROPOSAL APPROVAL FORM**

\*Write in Capital Letters Only

| **TITLE:** | DATA DEDUPLICATION | |
| --- | --- | --- |
| **THEME:** | AI POWERED SOLUTION FOR HUGE DATASETS | |
| **DOMAIN:** | MACHINE LEARNING | |
| **TYPE OF THE PROJECT:** | * Application * Product | * Research * Service |

**SUBMITTED BY:**

| **S.No** | **Name** | **USN** | **Signature** |
| --- | --- | --- | --- |
| 1 | EXSON NAVEIL PEREIRA | 4JK19CS017 |  |
| 2 | SHRADHA SAJEEVAN | 4JK19CS027 |  |
| 3 | SHRIJA B SHETTY | 4JK19CS028 |  |
| 4 |  |  |  |

**DEPARTMENT EVALUATION COMMITTEE**

| **S.No** | **Name** | **Approved/**  **Not Approved** | **Remarks** | **Signature** |
| --- | --- | --- | --- | --- |
| 1 |  |  |  |  |
| 2 |  |  |
| 3 |  |  |

**GUIDE ALLOTTED:**

| **S.No** | **Name** | **Designation** | **Domain Area** | **Signature** |
| --- | --- | --- | --- | --- |
| 1. |  |  |  |  |

**PO, PSO Relevance and Justification**

**\*** To be Filled by the Guide and Project Coordinator

| **S.No** | **PO, PSO** | | | **Justification** |
| --- | --- | --- | --- | --- |
| **1** |  | **PO1** | Engineering Knowledge |  |
| **2** |  | **PO2** | Problem Analysis |  |
| **3** |  | **PO3** | Design/Development of Solution |  |
| **4** |  | **PO4** | Conduct Investigation of Complex Problems |  |
| **5** |  | **PO5** | Modern Tool Usage |  |
| **6** |  | **PO6** | The Engineer and Society |  |
| **7** |  | **PO7** | Environment and Sustainability |  |
| **8** |  | **PO8** | Ethics |  |
| **9** |  | **PO9** | Individual and Team Work |  |
| **10** |  | **PO10** | Communication |  |
| **11** |  | **PO11** | Project Management and Finance |  |
| **12** |  | **PO12** | Life Long Learning |  |
| **13** |  | **PSO1** | Professional Skills |  |
| **14** |  | **PSO2** | Computational Skills |  |

| **Problem Statement:**  The problem is to determine how to design secure deduplication systems with higher reliability in cloud computing. Hence it is been proposed in the distributed cloud storage servers into deduplication systems to provide better fault tolerance. To protect data confidentiality, the secret sharing technique is utilized, which is also compatible with the distributed storage systems. To support deduplication, a short cryptographic hash value of the content will also be computed and sent to each storage server as the fingerprint of the fragment stored at each server |
| --- |
| **Scope/Objectives of the Project:**  Secure auditing and deduplication is big problem in cloud environment. This technique can be used for securely monitoring server space allocation. In this technique how to maintain back up data and remove unwanted file on server. |
| **Methodology of work: (Including diagram, flow chart and design calculations)**  It has shown how to design secure deduplication systems with higher reliability in cloud computing. By introducing the distributed cloud storage servers into deduplication systems to provide better fault tolerance. To further protect data confidentiality, the secret sharing technique is utilized, which is also compatible with the distributed storage systems. In more details, a file is first split and encoded into fragments by using the technique of secret sharing, instead of encryption mechanisms. These shares will be distributed across multiple independent storage servers. |
| **Expected Outcome of the Project:**  Deduplication can be run as an inline process as the data is being written into the storage system and/or as a background process to eliminate duplicates after the data is written to disk. |
| **Application of the Project:**  Data deduplication is a process that eliminates excessive copies of data and significantly decreases storage capacity requirements.  Deduplication can be run as an inline process as the data is being written into the storage system and/or as a background process to eliminate duplicates after the data is written to disk.  At NetApp, deduplication is a zero data-loss technology that is run both as an inline process and as a background process to maximize savings. It is run opportunistically as an inline process so that it doesn’t interfere with client operations, and it is run comprehensively in the background to maximize savings. Deduplication is turned on by default, and the system automatically runs it on all volumes and aggregates without any manual intervention.  The performance overhead is minimal for deduplication operations, because it runs in a dedicated efficiency domain that is separate from the client read/write domain. It runs behind the scenes, regardless of what application is run or how the data is being accessed (NAS or [SAN](https://www.netapp.com/data-storage/what-is-san-storage-area-network)).  Deduplication savings are maintained as data moves around – when the data is replicated to a DR site, when it’s backed up to a vault, or when it moves between on premises, [hybrid cloud](https://www.netapp.com/hybrid-cloud/what-is-hybrid-cloud), and/or public cloud. |
| **Advantages/Disadvantages:**  **Advantages:**  Greater backup capacityThe data is continuously validatedHigher data recovery service levelFacilitate the realization of backup data disaster recovery **Disadvantages:**   * With deduplication removing redundancies from a system, organizations using the technology could run an increased [risk of data corruption](https://social.technet.microsoft.com/wiki/contents/articles/31038.troubleshooting-data-deduplication-corruptions.aspx) if something goes wrong. For example, if the referenced data goes bad, then all of the data that points to it in your system goes bad too. Therefore, it's even more important to create backups of your data when using deduplication, just in case. |
| **Is the project proposed relevant to the Industry / Society or Institution?**  **Yes/No:**  Yes |
| **Can the product or process developed in the project be taken up for filing a Patent?**  **Yes / No:**  No |
| **Any other Technical Details (Please specify):** |

**Project Coordinator HoD CSE**